PRAXIS: Towards automatic cognitive assessment using gesture recognition

Farhood Negin, Pau Rodriguez, Michal Koperski, Adlen Kerboua, Jordi Gonzàlez, Jérémy Bourgeois, Emmanuelle Chapoulie, Philippe Robert, François Bremond

PRAXIS: Towards Automatic Cognitive Assessment Using Gesture Recognition

Farhood Negin[a,*], Pau Rodriguez[b], Michal Koperski[a], Adlen Kerboua[c], Jordi Gonzàlez[b], Jeremy Bourgeois[d], Emmanuelle Chapoulie[d], Philippe Robert[d], Francois Bremond[a]

[a]*STARS team - INRIA Sophia Antipolis, 06902 Valbonne, France*
[b]*Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Catalonia Spain*
[c]*Computer Science Department, College of NTIC, University of Constantine 2 - Abdelhamid Mehri, 25000 Constantine, Algeria*
[d]*the Cognition, Behaviour & Technology Unit (CoBTeK AI) and the CHU memory center at University Cote d'Azur. Institute Claude Pompidou, 10 rue Moliere, 06100 Nice, France*

Abstract

Praxis test is a gesture-based diagnostic test which has been accepted as diagnostically indicative of cortical pathologies such as Alzheimer's disease. Despite being simple, this test is oftentimes skipped by the clinicians. In this paper, we propose a novel framework to investigate the potential of *static* and *dynamic* upper-body gestures based on the Praxis test and their potential in a medical framework to automatize the test procedures for computer-assisted cognitive assessment of older adults.

In order to carry out gesture recognition as well as correctness assessment of the performances we have recolected a novel challenging RGB-D gesture video dataset recorded by Kinect v2, which contains 29 specific gestures suggested by clinicians and recorded from both experts and patients performing the gesture set. Moreover, we propose a framework to learn the dynamics of upper-body gestures, considering the videos as sequences of short-term clips of gestures. Our approach first uses body part detection to extract image patches surrounding the hands and then, by means of a fine-tuned convolutional neural network (CNN) model, it learns deep hand features which are then linked to a long short-term memory to capture the temporal dependencies between video frames.

*Corresponding author
Email address: farhood.negin@inria.fr (Farhood Negin)

We report the results of four developed methods using different modalities. The experiments show effectiveness of our deep learning based approach in gesture recognition and performance assessment tasks. Satisfaction of clinicians from the assessment reports indicates the impact of framework corresponding to the diagnosis.
Keywords: Human computer interaction, Computer assisted diagnosis, cybercare industry applications, human factors engineering in medicine and biology, medical services, monitoring, patient monitoring computers and information processing, pattern recognition.

---

1. Introduction

With overwhelming increase of computers in society and their ubiquitous influence in our daily activities, facilitating human computer interactions has become one of the main challenges in recent years. Hence, there has been a growing interest among the researchers to develop new approaches and better technologies to overcome this problem. The ultimate aim in this process is to achieve more sensor accuracy and efficiency of methods to bridge human-computer interaction gap and make it as natural as human-human interactions. Such methods will have a broad range of applicability in all aspects of life in a modern society from gaming and robotics to medical diagnosis and rehabilitation tasks. Considering recent progress of computer vision field, there has been an increasing urge upon medical domain. Computer-aided rehabilitation technologies are therefore gaining popularity among medical fraternity and are targeting more healthcare applications [1]. Employing Gesture recognition where human-computer interaction is indispensable, becomes one of the most favorable applications owing to its natural and intuitive quality.

Cognitive disorders such as Alzheimer's disease (AD) are prevalent among older adults. Studies show a maximum correlation between AD and limb apraxia in all phases of the disease [2]. One of the effective tests which has been developed to diagnose these disorders is the Praxis test. Praxis is defined as the ability to plan and perform skilled movements in a non-paralytic limb based on the previously learned complex representations. Accordingly, limb apraxia is inability to carry out a learned

motor act on command while there is no motor or sensory deficit in the subject [2, 3]. According to Geshwind's "disconnection model", apraxia is considered as failure (spatial or temporal error or failing to respond) of a subject to respond correctly with the limbs to a verbal command or having difficulty to imitate an action after being performed by an examiner [4]. Based on the American Psychiatric Association's report, Praxis test is accepted as diagnostically indicative sign of cortical pathologies such as AD [5]. However, the test is frequently neglected by clinicians despite being uncomplicated, straightforward and reliable estimate of the AD [6].

To capture changes in elderlies' behavioral pattern and to classify their cognitive status (Alzheimers disease - AD, mild cognitive impairment - MCI, healthy control - HC), there has been a lot of studies on patient monitoring and surveillance [7, 8, 9, 10] with a main focus on recognition of activities of daily living (ADLs) [11, 12]. The main goal of such frameworks is mostly to provide cost-efficient solutions for in-home or nursing homes monitoring. These systems try to alert the healthcare providers about a significant change in the ADL behavior pattern which may lead to cognitive impairment, falling of the patient or other health related changes. However, ADLs usually have a complex and highly-variable structure and need to be evaluated for a long period of time so as to be useful for clinicians to timely detect health deterioration in subjects.

Meanwhile, contact-based and various sensors for rehabilitation tasks [13, 14] have been developed and found practical applications such as post stroke recovery [15] and limb rehabilitation [16]. Having their own advantages and disadvantages, they have been mostly utilized in rehabilitation and not for assessment and diagnosis. The most prevailed field which has been applied for computer-assisted diagnosis is image processing. Machine learning algorithms fed with X-Ray, CT scan, MRI, retina images, *etc.*, which are de-noised, segmented, and represented, assist the clinicians with diagnosis or surgical planning through finding meaningful patterns [17]. While these methods provide valuable diagnostic information for surgical purposes, their need to use advanced hardware and to process huge datasets, which result in high cost for image interpretation, is a big drawback compared to cost-effective gesture recognition tasks. However, using gesture recognition to obtain an objective classification of a

Figure 1: The collected dataset consists of selected gestures for Praxis test. There are two types of gestures in the dataset: dynamic (14 gestures) and static (15 gestures) gestures. The dynamics are the ones including movement during the time that gestures are performed. The dynamic gestures are indicated with red arrows indicating their motion direction. On the other hand, the static gestures include body part orientation and position configuration without any movement during an amount of time. In another taxonomy the gestures are divided to: Abstract, Symbolic and Pantomimes (starting with "A", "S" and "P" respectively).

person's performance, particularly for medical diagnosis, still remains as a novel and largely unaddressed challenge for the research community.

Regarding the above-mentioned discussions, we have proposed a gesture recognition method by paying special attention to the Praxis test. The aim is to develop a robust and efficient computer-vision-assisted method to automatize the test procedure and to carry out assessments that help clinicians to have a more reliable diagnosis by providing a detailed analysis of subjects performances. Consequently, we have collected a challenging dataset [1] composed of dynamic and static gestures provided by clinicians for the Praxis test (Figure 1). We also adopt a gesture recognition framework, using a deep convolutional neural network (CNN) [18] coupled with a Longshort-term-memory (LSTM) [19], that jointly performs gesture classification and fine grained gesture correctness evaluation. As a result, we report performance of the proposed method and comparisons with developed baselines. With the evaluations we provide strong evidence about superiority of our representation learning method over traditional approaches, ensuring that robust and reliable assessments are feasible.

The remainder of this paper is organized as follows. In section II, we review the

[1] https://team.inria.fr/stars/praxis-dataset/

related studies on gesture recognition and computer-assisted rehabilitation and diagnosis. Section III introduces the formulation of our baseline methods and suggested CNN+LSTM model followed by section IV that presents the experimental analysis, results and discussions. Finally, section V concludes the study and discusses about future work.

2. Related Work

Contact based hand gesture or upper limb pose rehabilitation technologies are already in use in hospital and in-house environments with acceptable accuracy. However, design of these technologies comes with certain advantages and obvious limitations [20, 21]. For example, pattern recognition based prosthesis upper limb control in [22] obtained good results in controlled lab settings but it did not achieve anticipated results when it was tested in clinical real-world settings. While contact based systems achieved viable accuracy in different studies, their acceptability among users became restrained because of their dependency on experienced users. In order to be beneficial, the user needs to get accustomed to such devices. Being uncomfortable or even posing a health hazard are other disadvantages of these devices, as those are in physical contact with the users [23]. Because of their physical contact, mechanical sensor materials cause symptoms such as allergic skin reactions.

Other similar systems that have benefited from various modalities were also developed targeting full or body part rehabilitation [24]. Even virtual reality based methods have been tried for rehabilitation to recover patients from different disorders like for phantom limb pain [25] or recovering from chronic pain using serious gaming [26]. In a recent work [16], authors use a Leap motion sensor equipped with a gesture recognition algorithm to facilitate palm and finger rehabilitation. There are also other approaches which have been proposed in various domains but potentially can be adapted for rehabilitation and diagnosis contexts. For example [27, 28] try to evaluate choreography movements based on a gold-standard obtained from professional dancers. There are also lots of work that address the sign language recognition problem [29, 30, 31], where it may also require accurate reconstruction of hand shape. The challenge is to

match the gestures with corresponding words and construct conforming sentences.

Recently human action recognition has drawn interest among computer vision researchers due to its potential to improve accuracy of video content analysis [32, 33, 34, 35]. Although vision based systems are more challenging to develop and complex in configuration, they are more favorable in long term because of their user-friendly nature. Previously, most of the vision-based action recognition were based on sparse or dense extraction of spatial or spatio-temporal hand-crafted features [36, 37, 38, 39]. These methods usually consist of a feature detection and extraction step followed by a feature encoding step. For feature detection the most popular methods are Harris3D [40] and Hessian3D [41] while, for feature description HOG-HOF [40], HOG3D [42] and extended version of SURF descriptor [41] have found popularity in recent years. The most famous descriptor in recent times is improved dense trajectories [33] which reached state-of-the-art result on various datasets. However, it turned out that most of these methods are dataset-dependent and there is no all-embracing method that surpasses all the others [43]. Consequently, there is a growing interest in learning low- and mid-level features either in supervised or unsupervised ways.

Skeleton-based gesture and action recognition approaches have received lots of attention due to the immense popularity of Kinect-like sensors and their capability in body part detection. In many works [44, 45, 46, 47, 48], using skeleton and RGB-D cameras have shown advantages over methods using RGB videos by providing novel representation and well-crafted algorithms. The main challenges in skeleton-based methods other than noisy joint information and the occlusion problem are to deal with the high variability of gestures and movements, high dimensionality of the input and having different resolutions in temporal dimension (variable speed of gestures). Generally skeleton-based action recognition methods treat actions as a time series problem where body posture characteristics and dynamic of movements over time represent the actions [49]. A common approach for modeling the temporal dynamic of actions is using Hidden Markov Models (HMMs) or Temporal Pyramid models [50, 51]. While TP methods are restricted by the temporal windows size, HMMs face difficulty in finding the optimal temporal alignment of the sequences and the generative distribution in modeling long term contextual dependencies.

Late advancements in hardware development –particularly powerful GPUs– have been important in the revival of deep learning methods. Convolutional neural network architectures have become an effective tool for extracting high-level features and shown outstanding success in classification tasks [52, 53]. Recently, deep networks have also been adapted for hand [54, 55, 56] and body [57, 58] pose estimation and also gesture segmentation and recognition [59], achieving state-of-the-art results on ChaLearn gesture spotting challenge and also other challenging datasets. However, unconstrained training of complex neural network models requires a big amount of data. The most popular approaches to restrain the complexity of the model is to reduce the dimensionality of the input by applying smaller patch sizes or training the model in an unsupervised fashion [60, 61]. Conventional Recurrent Neural Network (RNNs) have also proved to learn the complex temporal dynamics of sequential data, first by mapping the data to a sequence of hidden layers, and then connect the hidden layers to outputs. Although RNNs have shown efficiency on speech recognition and text generation tasks, it has been shown that they have difficulty to learn long-term dynamics due to vanishing gradient problem. LSTMs provided a solution for this issue by allowing the model to keep information in hidden layer when it is necessary and update the layers when it is required. Since LSTMs are not confined to fixed length inputs or outputs they are practical for gesture recognition from video sequences and have shown success when unified with CNN features [62, 63, 64]. In this work, in order to avoid difficulties of temporal alignment in HMMs and learning long temporal dependencies in RNNs, we use LSTMs for modeling long temporal dependencies of the gesture sequences. Differently from [62, 63], we don't use 3D convolutions nor we train the CNN and LSTM jointly to adapt to the low hardware profile of hospital computers. Thus our approach resemble most to [64], although, differently from the latter, we design our pipeline to receive hand patches instead of whole images and perform feature fusion. This makes our model even more memory efficient than the previous ones since hand patches are much smaller than the whole scenes. In [64], regression is performed over pain scores. Differently, since we want to detect few incorrect frames in very long sequences, we face a highly imbalanced classification task for which we choose a weighted classification loss function.

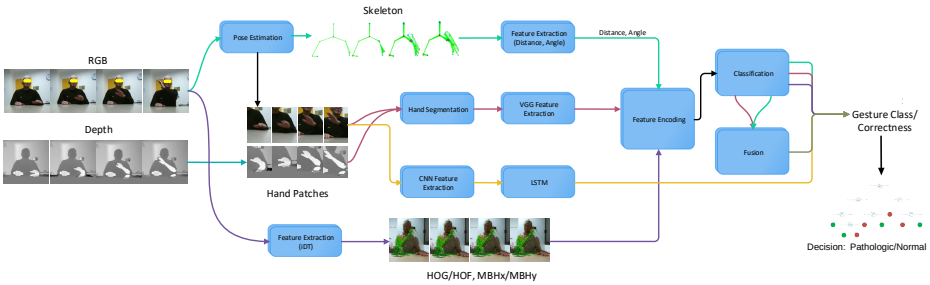


Figure 2: The data flow for the four method applied on the Praxis dataset. Flow of each method is separated by using a different color code.

3. Methodology

Next, we will define four methods we have applied to evaluate the dataset (Figure 2). Each path (indicated with different colors) learns its representation and performs gesture recognition independently given RGB-D stream and pose information as input.

Skeleton Based Method: Similar to [65] the joint angle and distance features are used to define global appearance of the poses. Prior to the classification (different from [65]), a temporal window based method is employed to capture temporal dependencies among consecutive frames and to differentiate pose instances by notion of temporal proximity.

Multi-modal Fusion: The skeleton feature captures only global appearance of a person, while deep VGG features extracted from RGB video stream acquire additional information about hand shape and dynamics of the hand motion which is important for discriminating gestures, specially the ones with similar poses. Due to sub-optimal performance of immediate concatenation of the high-dimensional features, a late fusion scheme for class probabilities is adopted.

Local Descriptor Based Method: Similar to action recognition techniques which use improved dense trajectories [35], a feature extraction step is followed by a fisher vector based encoding scheme.

Deep Learning based Method: Influenced by recent advancements in representation learning methods, a convolutional neural network based representation of hands is

coupled with a LSTM to effectively learn both temporal dependencies and dynamics of the hand gestures. In order to make decisions about condition of a subject (normal vs pathologic) and perform a diagnostic prediction, a decision tree is trained by taking output of gesture recognition task into account.

It should be noticed that for all of the developed methods we assumed that the subjects are in a sitting position in front of the camera where only upper-body of them are visible. We also assume that the gestures are already localized and the input to the system is short-term clipped videos. In the following sub-sections, we explain each method in more details.

3.1. Articulated Pose Based Action Recognition

Current depth sensors provide 25 or fewer articulated skeleton joints through their associated middleware including 3D coordinates on an axis aligned with the depth sensor. However, in near-range applications where accurate joint information is required, whenever optimal range of the sensor was not respected, the joints could get missed or mis-detected or the extracted information is noisy. Given our task, most of the time almost half of the subject's body is occluded and the subjects are very close to the sensor and some body parts get even closer during performing of the gestures. This leads to missing or noisy part detections by the sensor. Instead of using unreliable joint information, we use CNN-based body part detector from RGB images in [66] which returns 14 body parts. For our purpose only 8 upper body part joints are relevant ($N_j = 8$): *right hand*, *right elbow*, *right shoulder*, *left shoulder*, *left elbow*, *left hand*, *chin* and *top of the head*.

We formulate a pose descriptor similar to [65]. Following them, first, we calculate pairwise joint distances and angles at each frame and then, to augment the characteristics of the final descriptor we describe spatial and temporal relations between consecutive poses similar to [67, 68].

We represent the skeleton as a tree structure where the chin node is considered as the root node. The joint coordinates are transformed according to the root coordinate in order to eliminate the influence of joint positions with respect to the sensor coordinates. Before representation, to reduce jitter in estimated joints trajectories we

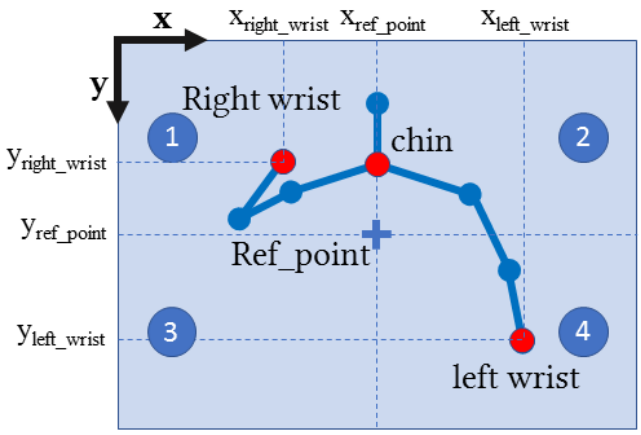


Figure 3: Dividing joint coordinates into four regions to detect the dominant hand in gesture performance

smooth joints position over temporal dimension by applying polynomial regression using weighted linear least squares and second degree polynomial model. Each subject performs similar gestures with variable speed resulting in variable frame sizes and joint trajectories. To achieve uniform performance speed along temporal dimension and to remove outliers in joints trajectories, once the smoothed joint positions are obtained, cubic interpolation of the values at neighboring joints is applied in the respective dimensions. Furthermore, to remove abrupt movements of the hand and elbow joints that are neither part of the gesture nor a jitter, a threshold is set which results in more stable joint values. Additionally, for the gestures in which laterality is not important (the subject is free to perform the gesture with either hand), we assume right hand as the dominant hand (considering that most of the subjects are right-handed) to reduce intra-class variability. Therefore, in these class of gestures, we mirror the instances performed by left hand according to a vertical line through a reference point defined as:

$$ref\_point = \left[x_{chin}, (y_{chin} + (y_{rhand} + y_{lhand})/_2)/2\right] \quad (1)$$

To find the gestures performed by left hand, we divide the skeleton's coordinate into four regions by setting the center to the calculated reference point (Figure 3). Having the joint trajectories, we can decide handedness of the performed gesture. Moreover, to compensate variations in body size, shape and proportions, we follow method in [69].

Starting from the root node (chin), we iteratively normalize body segments between the joints to average bone size in the training data.

To represent the skeleton, both joints' Euclidian distances and angles in polar coordinate are calculated using normalized joint positions. In order to preserve temporal information in pose representation, a feature extraction scheme based on temporal sliding window is adopted. At each time instance, Euclidian distances between all the joints are calculated. Besides, for each joint, distances from other instances' joints included in the sliding window is calculated and stored as well. If J_i^t represents features of joint i at time t and w shows the sliding window size: $J_i^t = [x_i^t, y_i^t]$ defines raw skeleton features at time t, where $i = 1, ..., 8$. Then, F^d calculates the distance descriptor:

$$F^d = \sqrt{(x_i^t - x_j^{t'})^2 + (y_i^t - y_j^{t'})^2} \tag{2}$$

Similarly, to calculate angular feature in polar coordinate we use:

$$F^a = \arctan\left(x_i^t - x_j^{t'}, y_i^t - y_j^{t'}\right) \tag{3}$$

where $t' \in \{t, t-1, ..., t-w\}, t' > 0$ and $i, j = 1, 2, ..., 8$ for both Eqs. 2 and 3.

Combining these features together, produces the final descriptor vector $F = [F^d, F^a]$ of dimension $N_f = 2 * w * {N_j}^2 = 1280$. To eliminate redundant information, PCA is applied on the position of torso joints and 512 dominant values preserving 99% of the descriptor information are kept. The final vector is normalized to zero mean and unit variance. The two feature types that capture dynamic of the gestures using sliding window produce some redundancy since several instances of the same frame are included in formulation of pose descriptor. While theoretically nonessential, this can be useful for classes with limited number of instances in the training data.

3.2. Multi-Modal Fusion

Skeleton-based descriptors have shown good classification accuracy for action recognition tasks where entire body is involved in performing the actions. In case of our problem, other than relative body part positions and orientations, detailed hand pose and finger articulation are also essential for recognition task. Since skeleton joints do

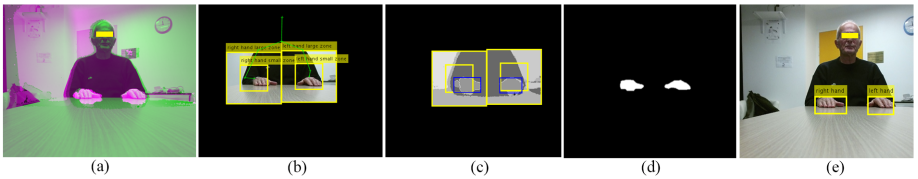


Figure 4: The steps of multi modal representation and recognition a) Registering depth image to align with RGB image b) Cropping the hand patches c) Clustering the depth values and detecting maximum overlap with the small patches d) Depth segmented hand blobs e) Register back accurate segmented hand blob on the RGB image and calculate bounding-box to extract image descriptors and fuse it with skeleton features.

not provide such detailed information, most of the gestures that can only be differentiated knowing subtle hand shape differences will not be recognized by a model that only relies on crude spatial information. We exploit depth data stream along with RGB images, first, to segment hand from the rest of body parts and then, to retrieve highly representative features only from the bounding-boxes surrounding the segmented hand (Figure 4).

Since working directly with input image and depth data from Kinect is computationally demanding, we use cropped patches around hands using skeleton joint information. First of all, using the depth and RGB camera intrinsics and their extrinsic relation, the depth data are registered on RGB images. Having depth and RGB registered, the hand skeleton joint is used for cropping the patches from the depth images. Accordingly, one big (160×160 pixels) and one smaller (80×80 pixels) square patches around the hand joints are cropped. For the depth images we only take the bigger patches which are Z-normalized. Later, we cluster the gray-level values in depth patches (to obtain hand blobs) using multi-level image thresholding by Otsu's method [70] which obtains the thresholds based on the aggregated histograms to quantize images. To detect the blob which most likely is the hand blob, we calculate the overlapping ratio of the blobs with the small patches' regions. The blob with the maximum overlap is selected as the hand blob. Finally, this hand blob is used to define the segmented hand bounding-box in RGB images.

Since CNNs have shown impressive results on various classification tasks, instead of hand-crafted image features, we use a pre-trained CNN model [71] (VGG-19) which

is trained on a subset of the ImageNet [72] database to extract deep features from the retrieved RGB bounding-boxes. The model is trained on more than a million of images on a wide range of image classes (1000 classes). There are 19 layers to learn weights from which 16 are convolutional layers and 3 are fully connected layers. To extract features, we use the patches as input to activate the convolutional layers and collect the features from the fully collected layer 'fc7' of size 4096 for each image patch.

Fusion: To combine the two modalities (skeleton+VGG image features) we follow a late fusion scheme by applying a simple linear combination of the obtained probabilities in the classification phase. If F is the final feature vector of the given video sequence v, $p(l_v|F)$ gives the probability of the predicted label l_v for that sequence and is calculated as:

$$p(l_v|F) \propto \alpha \cdot p(l_s|F^s) + (1-\alpha) \cdot p(l_d|F^d) \tag{4}$$

where l_s and l_d are predicted labels of the given video and $p(l_s|F^s)$, $p(l_d|F^d)$ are the probabilities of the skeleton and deep image patch descriptor modalities respectively. The coefficient α controls each modality's contribution which is set to 0.5 (through cross validation) indicating equal importance of the two modalities.

3.3. Descriptor Based Action Recognition

3.3.1. Action Descriptor Extraction

We use improved dense trajectories (iDT) [35] to extract local spatio-temporal descriptors. Dense trajectories ensure coverage of whole dynamic of the gestures which results extraction of meaningful features. Length of trajectories are limited to $t = 5$ frames to capture slight motion in consecutive frames. Short trajectories are more reliable than long ones, specially when there is a gesture with fast irregular motion or when the trajectories are drifting. Moreover, short trajectories are suitable for short term gestures like the ones available in our dataset. Similar to [35], we choose a space-time volume (i.e. patch) of size $S \times S$ pixels and t frames around each trajectory. For each patch around the trajectories we compute the descriptor vector $\mathbb{X}$ consists of HOG/HOF and MBHx/MBHy local descriptors.

3.3.2. Action Representation

The calculated descriptors are employed to create action representations based on Fisher vectors [73, 74]. Accordingly, first and second order statistics of a distribution of the feature set $\mathbb{X}$ are used for encoding a video sequence. Generative Fisher vector model is formed to model the features and the gradient of their likelihood are computed according to the model parameters (λ), *i. e.* $\Delta_\lambda \log p(\mathbb{X}|\lambda)$. The way the set of features deviates from their average distribution is depicted through a parametric generative model. To improve the learned distribution to further fit the observed data, a soft visual vocabulary is obtained by fitting a M-centroid Gaussian Mixture Model (GMM) into the training features within the preliminary learning stage:

$$p(x_i|\lambda) = \sum_{j=1}^{M} w_j g(x_i|\mu_j, \Sigma_j), \tag{5}$$

$$\mathbf{s.t.} \quad \forall_j : w_j \geq 0, \quad \sum_{j=1}^{M} w_j = 1, \tag{6}$$

$$g(x_i|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{D/2}|\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)}, \tag{7}$$

where $x_i \in \mathbb{X}$ represents a D-dimensional feature vector, $\{g(x_i|\mu_j, \Sigma_j)\}_{j=1}^{M}$ are the component of Gaussian densities and $\lambda = \{w_j, \mu_j, \Sigma_j\}_{j=1}^{M}$ are the parameters of the model: Respectively, $w_j \in \mathbb{R}_+$ is the mixture weights, $\mu_j \in \mathbb{R}^D$ is the mean vector, and $\Sigma_j \in \mathbb{R}^{D\times D}$ is the positive definite covariance matrices of each Gaussian component. The parameters λ are found using the Expectation Maximization restricting the covariance of the distribution to be diagonal. The GMM parameters are assessed through random sampling of a subset of $100,000$ features from the training set where the number of Gaussians is considered to be $M = 128$. Initialization of the GMM is performed ten times to obtain high precision and accordingly to provide the lowest error pertinent to the codebook. We define the soft assignment of descriptor x_i to the Gaussian j as a posteriori probability $\gamma(j|x_i, \lambda)$ for component j:

$$\gamma(j|x_i, \lambda) = \frac{w_j g(x_i|\mu_j, \Sigma_j)}{\sum_{l=1}^{M} w_l g(x_i|\mu_l, \Sigma_l)}, \tag{8}$$

Thereafter, the gradients of the j-th component can be calculated with respect to μ and σ using the following derivations:

$$\begin{aligned} G^{\mathbb{X}}_{\mu,j} &= \frac{1}{N_x\sqrt{w_j}} \sum_{l=1}^{N_x} \gamma(j|x_l,\lambda) \left(\frac{x_l - \mu_j}{\sigma_j} \right), \\ G^{\mathbb{X}}_{\sigma,j} &= \frac{1}{N_x\sqrt{2w_j}} \sum_{l=1}^{N_x} \gamma(j|x_l,\lambda) \left(\frac{(x_l - \mu_j)^2}{\sigma_j^2} - 1 \right), \end{aligned} \tag{9}$$

where N_x is the cardinality of the set $\mathbb{X}$. Finally, a set of local descriptors $\mathbb{X}$ as a concatenation of partial derivatives is encoded as a function of the mean $G^{\mathbb{X}}_{\mu,j}$ and standard deviation $G^{\mathbb{X}}_{\sigma,j}$ parameters for all M components:

$$V = [G^{\mathbb{X}}_{\mu,1}, G^{\mathbb{X}}_{\sigma,1}, ..., G^{\mathbb{X}}_{\mu,M}, G^{\mathbb{X}}_{\sigma,M}]^T. \tag{10}$$

The dimension of the Fisher vector representation is $2DM$.

3.3.3. iDT Based Action Recognition

To perform action classification, linear Support Vector Machines is employed. There are a lot of studies in the literature that reported high efficiency of linear classifier and good results obtained with high dimensional video representations such as Fisher vectors. Given a set of n instance-label pairs $(\mathbf{x}_i, y_i)_{i=1..n}$, $\mathbf{x}_i \in \mathbb{R}^k$, $y_i \in \{-1, +1\}$, we solve the following unconstrained optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \xi(\mathbf{w}; \mathbf{x}_i, y_i), \tag{11}$$

where C is a penalty parameter ($C > 0$) and $\xi(\mathbf{w}; \mathbf{x}_i, y_i)$ is a loss function $\max(1 - y_i\mathbf{w}^T\mathbf{x}_i, 0)$, referred to as L1-SVM. We set the parameter C to $C = 200$ which provides good results on a subset of training samples across various datasets. For multi-class classification, we implement the one-vs-all strategy.

3.4. Deep Learning Based Method

Inspired by the recent advances on facial motion recognition [64], we propose to use a CNN to extract spatial static hand features, and learn their temporal variation by using Long Short-Term Memory (LSTM) [19]. Different from [64], the pipeline has

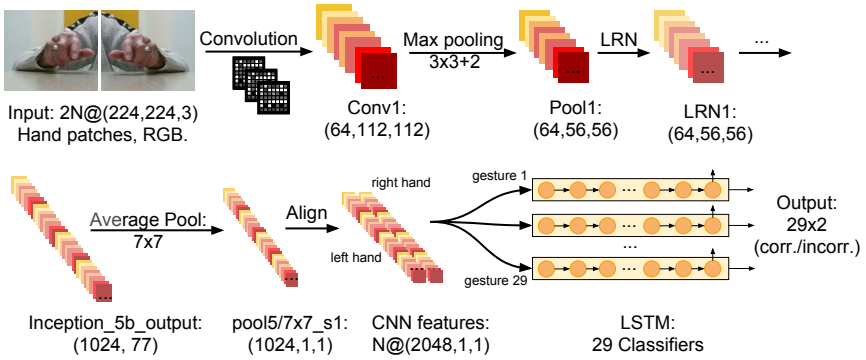


Figure 5: The proposed pipeline for hand configuration representation and gesture recognition. Spatial information is extracted from hand patches by feeding them to a CNN, and temporal information is leveraged using Long Short-Term Memory.

been modified so as to temporally align the patches from both hands, and the use of a weighted loss function so as to increase the sensitivity to incorrect gestures, which are important to detect. As it can be observed in Figure 5 the proposed pipeline is divided in three main stages: (i) hand patch extraction, (ii) CNN fine-tuning and feature extraction, and (iii) temporal aggregation with the LSTM. These three stages are next described in detail.

3.4.1. Hand patch extraction

Similar to the preprocessing steps in multi-modal method we extract body parts and using hand joints we extract image patches around both hands. In order to avoid the ambiguity in detecting the active hand, the same pre-processing step for flipping left and right hands in lateral gestures are also applied before sending the patches as input to the training network.

3.4.2. Hand Gesture CNN

In order to extract highly discriminative spatial features from the hand patches, we first fine-tune a CNN to classify the gesture and whether the gesture is correct or incorrect. For this purpose a GoogleNet architecture [75] is chosen since it has shown to provide competitive results while being lightweight compared to other models such as VGG [71]. Moreover following [76], we initialize the CNN with Deep Hand [77], a

GoogleNet model trained with Expectation Maximization (EM) on approximately one million images to predict 60 different gestures.

Concretely, we reinitialize all the weights in the loss streams of the GoogleNet, and fine-tune the network with the data presented in this work. In order to force the network to find highly discriminative features, the two output layers are reshaped to predict a probability distribution over 58 labels, where the first half corresponds to the 29 correctly-executed gestures, and the second half corresponds to their incorrect execution.

The hand gesture CNN is trained with Stochastic Gradient Descent (SGD) by minimizing the cross-entropy loss function using the Caffe Deep Learning Framework [78] during ten epochs, with a learning rate of 0.001 except for the reinitialized layers, for which is ten times higher. Standard data augmentation is performed by extracting random 224×224 sub-crops from the hand patches, and by randomly performing horizontal flips, *i.e.* randomly flipping the image crops along a central vertical axis following a *Bernoulli* distribution with $p = 0.5$.

After fine tuning, feature activation maps for the whole dataset are extracted from the last pooling layer. These feature vectors have a dimensionality of 1024. Once extracted, feature vectors from both hands in the same frame are concatenated, forming a 2048-dimensional feature vector. This concatenated vector is then fed to a LSTM, which will be explained next, in order to leverage the temporal information present in the videos to make the final prediction.

3.4.3. Aggregating temporal information

Given a set of consecutive frames $F = \{f_1, ...f_n\}$ we are interested in recognizing the gesture represented in those frames $p_g = p(gesture|F)$ and whether the gesture is correct or incorrect $p_c = p(correct|F)$. Hence, LSTMs are especially suited for this problem, since they are able to model long term dependencies by solving the problems of vanishing and exploding gradients through a series of gates [19] known as input, output, and forget gates, which regulate the flow of information in the LSTM cell.

Given the features of both hands extracted from the CNN that correspond to F, two independent LSTMs are trained by means of Backpropagation Through Time (BPTT)

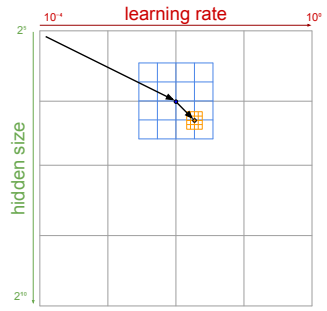


Figure 6: 2D gridsearch example. Best combinations are found iteratively from coarse to fine.

[79] so as to model p_c, and p_g respectively. Differently from [80], where the Mean Squared Error (MSE) is minimized on each frame, the LSTMs used in this work are trained to minimize the cross-entropy error of single predictions on whole video sequences, thus zeroing out the output and gradients of intermediate frames. In order to overcome the bias towards correct predictions due to the data imbalance, the loss function for p_c was weighted to increase the sensitivity to correct examples. Concretely, it was changed from:

$$loss(\mathbf{O}, c) = -\mathbf{O}_c + log(\sum_j e^{O[j]}), \tag{12}$$

where $\mathbf{O}$ is a 2-d vector containing p_c, and $c \in \{0, 1\}$ is the class label (incorrect, correct), to:

$$loss(\mathbf{O}, c) = (1 - p(c))(-o_c + log(\sum_j e^{o_j})). \tag{13}$$

Since $p(c)$ corresponds to the fraction of training video sequences labeled as c, and given that incorrect gesture sequences are underrepresented in the dataset, multiplying the loss by $1 - p(c)$ increases the penalty of misclassifying an incorrect gesture.

The LSTMs are trained with torch[2] using Adam [81] until they reach a plateau. Weights are initialized by sampling from a uniform distribution $unif\{-0.8, 0.8\}$, and the network architecture and hyperparameters are chosen by gridsearch, see Figure 6

[2] `torch.ch`

for an example.

In order to compare the diagnostic performance of LSTM classifier with clinician's decisions, a decision tree is trained using outcome of gesture correctness test. The best pruning level of the decision tree is calculated with cross validation method. Therefore, the correctness results of a subject performing the gestures are exposed to the decision tree and resulted in a decision whether a subject is normal or pathologic. Another decision tree is trained using ground-truth labels of gesture correctness test which is annotated by the clinicians. Comparison between the classification performance of the two decision trees interestingly shows how the LSTM classifier outperforms clinicians in diagnostic decisions based on a subject's performance which accordingly develops an objective criteria by global learning dynamics of the gestures in the whole dataset.

4. Experiments and analysis

4.1. Dataset

We collected a new challenging RGB-D upper-body gesture dataset recorded by Kinect v2. The dataset is unique in the sense that it addresses the Praxis test, however, it can be utilized to evaluate any other gesture recognition method. List of the gestures, their assigned ID and a short description about them is shown in table 1. Each video in the dataset contains all 29 gestures where each one is repeated for 2-3 times depending on the subject. If the subject performs the gesture correctly, based on decision of the clinician, the avatar continues the experiment with the next gesture, otherwise, they repeat it for 1-2 more times. Using the new Kinect v2 we recorded the videos with resolution of RGB: 960×540, depth: 512×424 without human skeletons information. The videos are recorded continuously for each subject. The dataset has a total length of about 830 minutes (with average of 12.7 minutes for each subject).

We ask 60 subjects to perform the gestures in the gesture set. From the subjects, 29 were elderly with normal cognitive functionality, 2 amnestic MCI, 7 unspecified MCI, 2 vascular dementia, 10 mixed dementia, 6 Alzheimer patients, 1 posterior cortical atrophy and 1 corticobasal degeneration. There are also 2 patients with severe cognitive impairment (SCI). We didn't use the two SCI patients' videos in the experiment since

Table 1: List of the available gestures in the dataset and corresponding information.

Category	Uni/Bimanual	ID	Type	Description	Similar gestures
Abstract	Unimanual	A1-1	Static	Left hand on left ear	A1-2, A1-3, A1-4, S1-1, S1-2, S1-5, P1-5
		A1-2	Static	Left hand on right ear	A1-1, A1-3, A1-4, S1-1, S1-2, S1-5, P1-5
		A1-3	Static	Right hand on right ear	A1-1, A1-2, A1-4, S1-1, S1-2, S1-5, P1-5
		A1-4	Static	Right hand on left ear	A1-1, A1-2, A1-3, S1-1, S1-2, S1-5, P1-5
		A1-5	Static	Index and baby finger on table	P1-3, P1-4, A2-2
	Bimanual	A2-1	Static	Stick together index and baby fingers	S2-1, S2-4, P2-1, A2-2, A2-5, A2-3, A2-4
		A2-2	Dynamic	Hands on table, twist toward body	P2-2, P1-4
		A2-3	Static	Bird	A2-1, A2-4, A2-5, S2-1, S2-4
		A2-4	Static	Diamond	A2-1, A2-3, A2-5, S2-1, S2-4
		A2-5	Static	ring together	A2-1, A2-3, A2-4, S2-1, S2-4
Symbolic	Unimanual	S1-1	Static	Do a military salute	A1-1, A1-2, A1-3, A1-4, S1-2, S1-4, P1-1, P1-3
		S1-2	Static	Ask for silence	A1-1, A1-2, A1-3, A1-4, S1-1, S1-4, P1-1, P1-3, P1-5, S1-3
		S1-3	Static	Show something smells bad	S1-2, S1-5, S2-4, P1-2, P1-5
		S1-4	Dynamic	Tell someone is crazy	P1-1, P1-3, A1-1, A1-2, A1-3, A1-4
		S1-5	Dynamic	Blow a kiss	S1-2, S1-3, P1-5
	Bimanual	S2-1	Dynamic	Twiddle your thumbs	S2-4, P2-1, A2-5
		S2-2	Static	Indicate there is unbearable noise	S2-3, S2-4, P2-4, P1-1
		S2-3	Static	Indicate you want to sleep	S2-2, S1-1, S2-4, A1-1, A1-2, A1-3, A1-4
		S2-4	Static	Pray	S1-2, S1-3, S1-5, S2-3, A2-5
Pantomime	Unimanual	P1-1	Dynamic	Comb hair	S1-1, S1-4, P1-3, A1-1, A1-2, A1-3, A1-4
		P1-2	Dynamic	Drink a glass of water	S1-2, S1-3, S1-5, P1-5
		P1-3	Dynamic	Answer the phone	P1-1, S1-1, S1-4, A1-1, A1-2, A1-3, A1-4
		P1-4	Dynamic	Pick up a needle	P2-1, P2-3
		P1-5	Dynamic	Smoke a cigarette	P1-2, S1-2, S1-3, S1-5
	Bimanual	P2-1	Dynamic	Unscrew a stopper	S2-1, P2-5, A2-5, P2-4
		P2-2	Dynamic	Play piano	P2-5, A2-2
		P2-3	Dynamic	Hammer a nail	P1-4, P2-5, P2-4
		P2-4	Dynamic	Tear up a paper	P2-3, P2-1, P2-5
		P2-5	Dynamic	Strike a match	P2-1, P2-3, P2-4

their performances were erratic and noisy and not useful for current study. However, we kept them in the dataset for further studies.

All of the videos are recorded in office environment with fixed position of the camera while subjects sit behind a table where only their upper body is visible. The dataset is composed of fully annotated 29 types of gesture (14 dynamic, 15 static). All of the gestures are recorded with fixed ordering, though the repetition of each gesture could be different. There is no time limitation for each gesture which makes the participants to finish their performance naturally. Laterality is important for some of the gestures. Therefore, if these gestures are performed with the opposite hand, those are labeled as "incorrect" by the clinician. A 3D animated avatar administrates the experiments (Figure 7). First, she starts with performing each gesture by precisely explaining how the participant should perform it. Next, she asks the participant to perform the gesture by sending a "Go" signal. The gestures are also divided into three main categories: Abstract, Symbolic and Pantomime gestures abbreviated by A, S, and P, respectively (Figure 1).

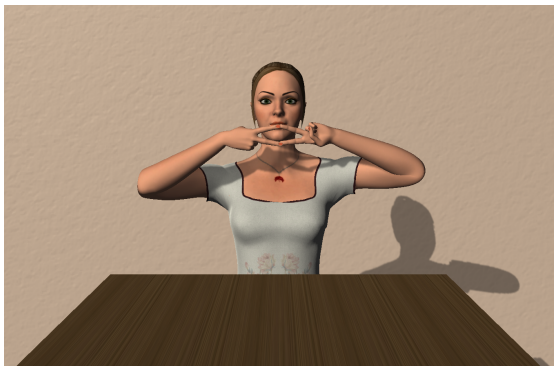


Figure 7: The virtual avatar guides the patients in a virtual environment.

Although the dataset was collected using the same setting for all of the subjects, it is still challenging because of the selected gestures and the subjects who are real cognitive patients coming to memory center. For some of the gestures in the dataset only hand pose differs but the whole body part configuration and gesture dynamics are very similar as shown in Figure 8.

The main focus in the dataset is on two tasks: ”gesture recognition” which consists in learning to recognize gestures from several instances of each category performed by different subjects and ”correctness of performance” which is the evaluation of gestures based on quality of performance by each subject. The second task is more challenging since the ”correctness” is subjective and depends on the professional opinion of the clinician and is not obvious all the times. The dataset will be made publicly available for research community to bring more contributions on this task.

For the experiments we follow three-folds cross validation protocol, in which we divide the dataset into three nearly balanced subsets (patients 1-16, 17-37, and 38-58) . At each fold we run the training with the videos in the current fold and we use the two other subsets for validation and monitoring of training performance and also hyper-parameters optimization and finally testing.

4.2. Results and Discussion

In this work we made a stride towards non-invasive detection of cognitive disorders by means of our novel dataset and an effective deep learning pipeline that takes into ac-

Figure 8: Examples of challenging cases in Praxis gesture dataset. Some of the gestures are very similar in upper-body and arm movement and only differs in hand pose (a) and (b). Almost half of the gestures require both hands to perform e.g. (c, g). Some dynamic gestures are very similar and just differ in speed and range (c, d). Performer variation in upper body dynamics: some of the subjects keep their upper-body steady, while the others aim toward the camera (g, h). For some other gestures, dynamic of the gesture differs totally from subject to subject where some subjects gesticulate more (e, f). In some gestures subtle hand movements make the difference between correct and incorrect performances which makes the recognition task very challenging (i, j, k, l).

count temporal variations, achieving 90% average accuracy on classifying gestures for diagnosis. The performance measurements of the applied algorithms are given in table 2. In both tasks (gesture and correctness classification) concatenated dense trajectory based local descriptors performs relatively better than the other baselines, specially, in dynamic gesture category. Particularly in gesture classification of dynamic gestures its performance is almost identical to CNN+LSTM approach. One possible explanation is that MBH descriptors are good in encoding motion pattern and since dynamic gestures include lots of motion they are capable of capturing them. They perform poorly in correctness of static gestures since 60 to 70 percent of frames in static gestures are static gestures do not contain any motion and the subject is in stable position in a spec-

Table 2: Comparison of the obtained results using proposed method in terms of accuracy of gesture classification and correctness of performance with other baseline methods.

		Accuracy			Correctness		
Method		Static	Dynamic	Average	Static	Dynamic	Average
Skeleton	Distance	70.04	56.99	63.51	72.04	59.93	65.98
	Angle	57.21	51.44	54.32	68.13	62.16	65.14
	Distance+Angle	61.83	55.78	58.80	70.06	61.49	65.77
Multimodal Fusion	RGB (VGG)	67.63	63.18	65.40	68.21	63.54	65.87
	RGB (VGG)+Skeleton	72.43	62.75	67.59	70.72	64.55	67.63
improved dense trajectories (iDT)	HOG/HOF	65.04	61.31	63.17	61.89	57.37	59.63
	MBHx/MBHy	70.32	75.49	72.90	55.63	72.93	64.28
Deep Learning	CNN+LSTM	**92.88**	**76.61**	**84.74**	**93.80**	**86.28**	**90.04**

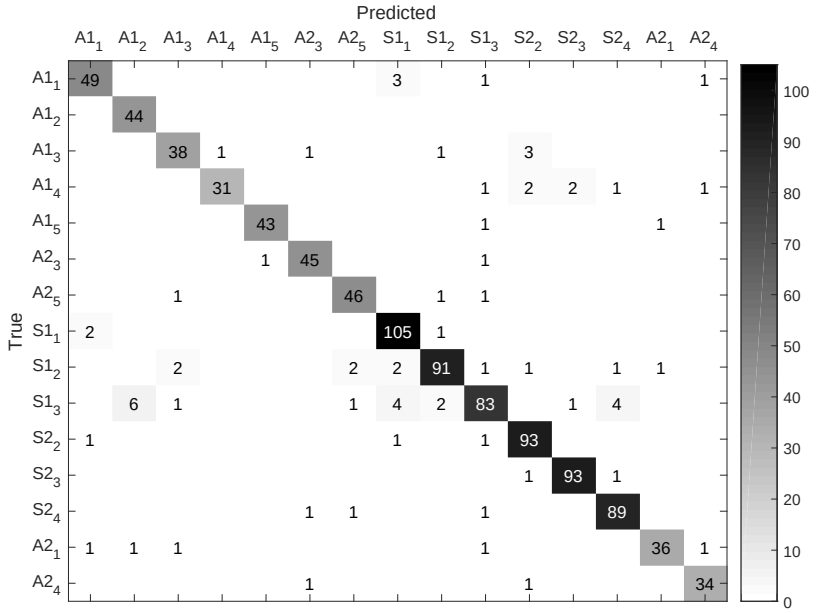
ified gesture's key frame. CNN+LSTM does not perform good in dynamic gestures as good as static one, possibly because of the high variation in dynamic gestures. It is interesting to see that, by using distance feature in articulated skeleton based approach, we obtain competitive results compared to the other baselines. We hypothesize that the good results are obtained due to the robust skeleton joint information and highly varied data in the dataset. However, this method performs poorly when it comes to dynamic gesture classification. The reason for its poor performance might be lack of enough articulation in hand poses when we solely rely on the joint information specially in the gestures which upper-body configuration does not differ between gestures (e.g. Fig. 8 e, f). The results also demonstrate that the combination of both modalities (skeleton with image patches) is more robust and reduces confusion as shown by increase in the recognition rate of gesture classification of static category and correctness of static and dynamic categories.

As can be observed the proposed method outperforms all the baselines in all of the tasks. It is important to note that these results are obtained by using gesture-wise LSTMs on hand patch data extracted from a CNN trained for classifying correctness and gesture simultaneously. Hence, since the task performed by the CNN was harder, it had to learn more discriminative features which then could be used by the LSTMs to better classify the video sequences. The existence of static and dynamic gestures

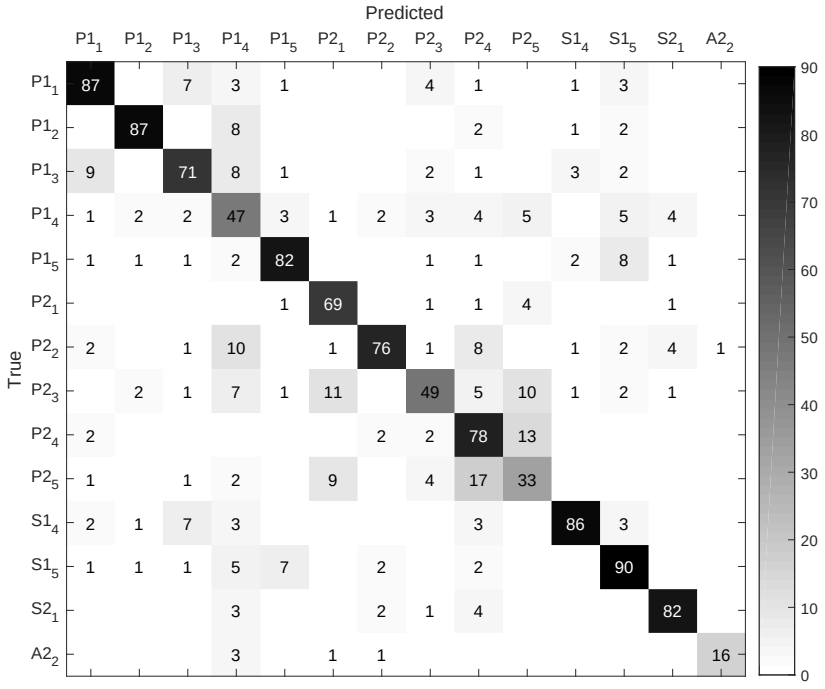
did also condition the decision of using individual LSTM classifiers since 1 layer and 32 hidden units sufficed for most of the static sequences while the dynamic sequences needed up to 6 layers and 256 hidden units. This was expected since LSTMs that classified dynamic gestures had to model complex temporal relationships while the static gesture LSTMs needed only to find the exact frame where the gesture was performed and apply a linear classifier on the frame CNN features. Additionally, the fact that the LSTMs were trained gesture-wise allowed us to use sequences from other similar gestures as negative samples during training. It is interesting to see how our representation learning method outperforms all of the hand-crafted feature methods' performance. It is unlikely that having more data will improve hand-crafted methods' performance. However, it is highly expected that as more training data become available, the representation learning approach will achieve even more accuracy and better suited for independent settings.

The confusion matrices in figure 9a and 9b illustrate the behavior of our CNN+LSTM method in gesture classification task. The superior performance of the classifier in static gestures classification is immediately apparent. It can be noticed that some gestures are easily classified. This is the case for gesture $A1\_2$ that is always classified correctly and its highest false positive (FP) belongs to the class $S1\_3$ whose arm configurations during the static frames are identical. In dynamic gestures there are more confusions which most of them are because of resemblance in body and arm configurations and also variations coming from performer that gesticulate more or does extra arbitrary motions. The clearest example of this confusion is between gesture $P2\_4$ and $P2\_5$ (figure 8) where the pantomime gesture "tearing a paper" is very similar to "lighting a match" gesture and the only difference to separate the two is the speed of performing the gesture.

From clinician point of view fine-grained gesture classification is not important. What concerns them is evaluation of gesture correctness. They already know which gesture the subject is asked to perform (class label) and what is important is to know if that specified gesture is carried out correctly or not. Tables 3 and 4 illustrate detailed gesture correctness evaluation at each fold on static and dynamic gestures respectively. For each gesture we achieve an acceptable accuracy that ensures robustness of the



(a) Static gestures



(b) Dynamic gestures

Figure 9: Confusion Matrices for the predicted gestures. The number in each element of the matrices indicates the number of predicted instances.

Table 3: Results in terms of correctness of performance for each fold in static gestures.

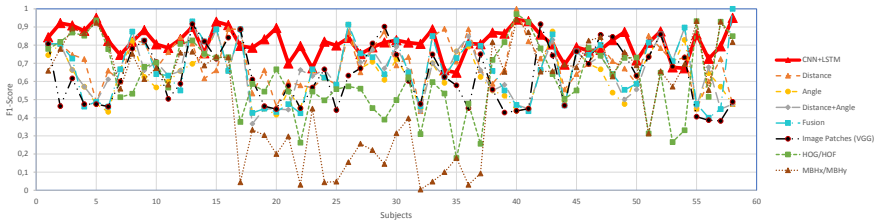
Static				
Gesture	Folds			
	1	2	3	Average
S1_1	1	0.952	1	0.984
S1_2	0.955	0.930	1	0.961
S1_3	0.906	0.925	1	0.943
S2_2	1	0.906	0.968	0.958
S2_3	0.978	1	1	0.992
S2_4	0.933	0.951	0.885	0.923
A1_1	1	1	1	1
A1_2	1	1	1	1
A1_3	0.968	1	1	0.989
A1_4	0.969	1	1	0.989
A1_5	0.903	0.900	1	0.934
A2_1	0.833	0.742	0.789	0.788
A2_3	0.870	0.851	0.900	0.874
A2_4	0.833	0.694	0.800	0.775
A2_5	0.923	0.920	1	0.947

Table 4: Results in terms of correctness of performance for each fold in dynamic gestures.

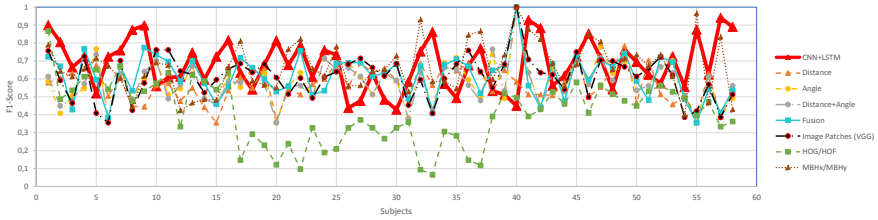
Dynamic				
Gesture	Folds			
	1	2	3	Average
S1_4	0.976	1	0.941	0.972
S1_5	0.891	1	1	0.963
S2_1	0.882	0.906	0.937	0.908
P1_1	0.895	0.854	0.968	0.906
P1_2	0.800	0.866	0.875	0.847
P1_3	0.730	0.888	0.937	0.852
P1_4	0.745	0.836	0.781	0.787
P1_5	0.869	0.880	0.968	0.906
P2_1	0.769	0.795	0.875	0.813
P2_2	0.857	0.906	1	0.921
P2_3	0.814	0.750	0.810	0.791
P2_4	0.869	0.880	0.777	0.842
P2_5	0.666	0.711	0.795	0.724
A2_2	0.846	0.794	0.880	0.840

classifier which is very important for diagnosis task. Again it immediately becomes evident that the performance in static gestures (12 out of 15 class's accuracy is higher than 90%) category surpass dynamic category, although, there are more instances of dynamic gestures in the dataset and intuitively it is more likely for the classifier to learn the dynamics of these gestures. But it seems that complexity of these categories and nuances of gesture correctness of some of the gestures are too much to be learned with available number of trials. This gives a hint for clinical aspect of the work that the static category is more appropriate one and should contribute more in later data collections and more gesture classes of this category should be included in order to have more reliable evaluations.

Capturing incorrect performances are of utmost importance that small nuance can affect accuracy of the diagnosis reports. This is because some gestures are simple enough for the subjects and most of the time are performed correctly while it is important and decisive to capture incorrect performances. This problem is rooted in unbalanced dataset where some classes have a few instances of incorrect performances. Although, the problem rectified somehow using similar gestures and employing the loss function, the nature of incorrect performances still remains undefined. Incorrect gestures could include anything and this makes these classes highly variable. Similar gestures stay far from real incorrect instances of a class and in some cases it might cause even more confusion. For example, we take gesture $P2\_2$ which is "playing piano" gesture as similar gesture for abstract gesture $A2\_2$ but in practice when a patient performs $P2\_2$ incorrectly, the incorrect performance is very close to $P2\_2$ and far from $A2\_2$. Moreover, in practice there are some subject specific redundant movements. For example, some subjects have specific mannerism and repeat it sporadically (one subject fixes his glasses before every performance and another one aims towards the examiners and asks questions). Although these subjects perform the gestures correctly but these additional movements hinder the proper evaluation. Ideally these subject specific movements could be learned and filtered out during pre-processing phase. In order to show the effectiveness of the proposed approach on evaluation of performance across individuals which is essential in terms of diagnosis, we conduct a comparative analysis using F1-score (figure 10). It can be observed that for most of the subjects CNN+LSTM



(a) Static gestures



(b) Dynamic gestures

Figure 10: The comparison of F1-scores with respect to subjects obtained by different methods for (a) static and (b) dynamic gestures. The proposed method (highlighted by red) shows better F1-score for most of the subjects and is less erratic compared to the others.

surpass the other methods acquiring higher F1-score underlying that CNN+LSTM is more consistent and reliable as compared to the other baselines specially when static gestures are taken into consideration. The highest F1-score fluctuations happen for subjects #15 to #40 where it can be verified that CNN+LSTM shows less fluctuations with an average score of 82% when compared to the others.

Finally, to delve deeper into the details of cognitive assessment of the subjects, we need to highlight the importance of the correctness classification of the gestures. As the classifier is only trained on correctness labels of the given instances, there is no immediate correlation between correctness of a gesture and condition of a subject. For example, a subject can perform one gesture correctly and the condition of the subject could be either normal or pathologic and therefore can not be inferred by relying on the correctness of that specific gesture. To ascertain the link between the correctness information of the gesture performances and the health status (Normal versus pathologic) of a subject, a pattern analysis needs to be carried out. Knowing knowledge discovery quality of decision trees and their high predictive performance, a tree model is trained given both overall performance of subjects on the gesture set and their condi-

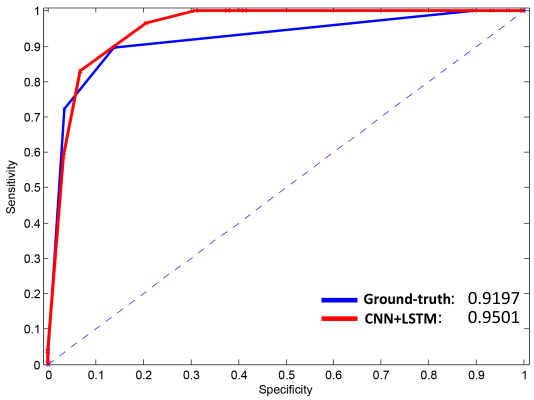
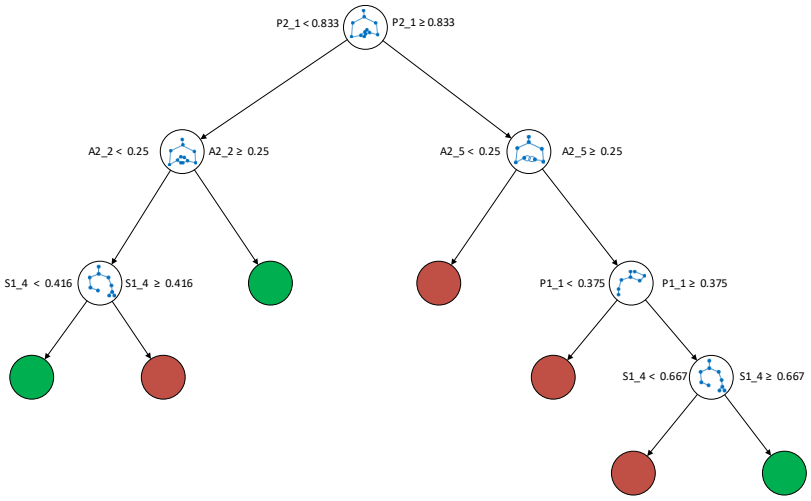
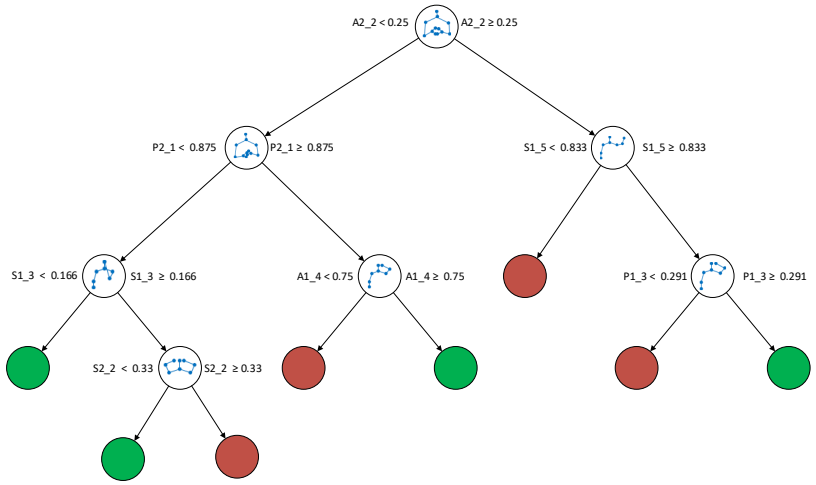


Figure 11: ROC of diagnostic classification using decision trees.

tion as input. $F = \{f_i | i = 1 \dots 29\}$ is the normalized feature vector of a subject where f_i belongs to a gesture in the dataset showing the performance of the subject on that gesture. To verify the efficacy of the predictions obtained by the LSTM classifier, two feature vectors are created for each subject; one from ground-truth correctness values (labeled by clinicians) and the other one using correctness labels produced by the classifier. Then, the decision tree is trained to predict the condition of the subject whether it is normal or pathologic. Figure 11 illustrates performance of the trained classifiers. Using the ground-truth labels, the decision tree can decide about condition of the subjects with 92% accuracy, whilst this rate is 95% when predictions related to the LSTM classifier are used. The accuracy difference of the two predictions (3%) is related to only two patients. The low rate of discrepancy between the ground-truth and classifier's diagnostic predictions encourages that the objective assessment is achievable when diagnostic-specific training is targeted. This also implies that all the diagnostic information can not be mined only observing the gestures and the clinicians subjective opinions play an important role in providing final diagnoses. The trained decision trees are depicted in figure 12. The most decisive gestures in diagnosis can be seen in nodes of the generated trees. Gestures $A2\_2$ and $P2\_1$ appear on root and first child node of both trees denoting their high impact contribution in diagnosis. Although it was observed that the accuracy of the classifications of the static gestures is higher than



(a) Ground-Truth



(b) CNN+LSTM

Figure 12: Resulted trees illustrated using the trained decision tree classifier. Green leaves represents ”Normal”, while red leaves indicates ”Pathologic” subject.

that in the dynamic gestures, the most important gestures appeared in the node of the trees belong to both categories (4 static and 6 dynamic). In total, there are 10 different gestures selected by the decision trees showing that an optimal subset of gestures and subsequently a shorter Praxis test consisted of lower number of gestures could be practiced. However, the trees are self-explanatory and very easy to follow and they are therefore comprehensible by the clinicians and even if it is required they can explain the performance of a subject and argue about the decision. Moreover, using the trees, a descriptive set of rules can be generated which explains what kind of performance would lead to an specific opinion. Further analysis can be carried out by applying different data mining techniques to interpret the results and this will be investigated in our future study.

5. Conclusion

Early diagnosis of cognitive impairments are essential to provide better treatment for elderlies. Praxis test is accepted as diagnostically indicative sign of cortical pathologies such as AD. Despite being uncomplicated, straightforward and reliable estimate of the AD, the test is frequently ignored by clinicians. To avoid such situations which arise during this process, we proposed a computer-assisted solution to undergo evaluation of automatic diagnosis process with help of computer vision. The evaluations of the system can be delivered to the clinicians for further assessment in decision making processes. We have collected a unique dataset from 60 subjects and 4 clinicians targeting analysis and recognition of the challenging gestures included in the Praxis test. To better evaluate the dataset we have applied different baseline methods using different modalities. Using CNN+LSTM we have shown strong evidence that complex near range gesture and upper body recognition tasks have potential to be employed in medical scenarios. In order to be practically useful, the system must be evaluated with larger population. However, satisfactory feedback of clinicians from our preliminary evaluations is a promising commencement.

Acknowledgment

The research leading to the results obtained in this work has been partially supported by the French ANR Safee project, INRIA Large-scale initiative action called PAL (Personally Assisted Living), the Spanish project TIN2015-65464-R (MINECO/FEDER), the 2016FI_B 01163 grant of Generalitat de Catalunya, and The European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307.

References

[1] J. Zariffa, J. D. Steeves, Computer vision-based classification of hand grip variations in neurorehabilitation, in: Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on, IEEE, 2011, pp. 1–4.

[2] S. R. Chandra, T. G. Issac, M. M. Abbas, Apraxias in neurodegenerative dementias, Indian journal of psychological medicine 37 (1) (2015) 42.

[3] R. L. Heilman KM, Apraxia, Clinical Neuropsychology 128 (10) (2003) 215–235.

[4] M. Catani, et al., The rises and falls of disconnection syndromes, Brain 128 (10) (2005) 2224–2239.

[5] A. P. Association, Diagnostic and statistical manual of mental disorders, text rev.).

[6] P. Peigneux, M. Van der Linden, D. Le Gall, Evaluation des apraxies gestuelles, L'apraxie, 2 (2003) 133–138.

[7] T. Banerjee, J. M. Keller, M. Popescu, M. Skubic, Recognizing complex instrumental activities of daily living using scene information and fuzzy logic, Computer Vision and Image Understanding 140 (2015) 68–82.

[8] D. Brulin, Y. Benezeth, E. Courtial, Posture recognition based on fuzzy logic for home monitoring of the elderly, IEEE transactions on information technology in biomedicine 16 (5) (2012) 974–982.

[9] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2847–2854.

[10] F. Negin, S. Cogar, F. Bremond, M. Koperski, Generating unsupervised models for online long-term daily living activity recognition, in: Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on, IEEE, 2015, pp. 186–190.

[11] K. Avgerinakis, A. Briassouli, I. Kompatsiaris, Recognition of activities of daily living for smart home environments, in: Intelligent Environments (IE), 2013 9th International Conference on, IEEE, 2013, pp. 173–180.

[12] A. König, C. F. Crispim-Junior, A. G. Uria, F. B. Covella, A. Derreumaux, G. Bensadoun, R. David, F. Verhey, P. Aalten, P. Robert, Ecological assessment of autonomy in instrumental activities of daily living in dementia patients by the means of an automatic video monitoring system, ICT for assessment and rehabilitation in Alzheimers disease and related disorders (2016) 29.

[13] C. W. Tan, S. W. Chin, W. X. Lim, Game-based human computer interaction using gesture recognition for rehabilitation, in: Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on, IEEE, 2013, pp. 344–349.

[14] L. E. Sucar, R. Luis, R. Leder, J. Hernández, I. Sánchez, Gesture therapy: A vision-based system for upper extremity stroke rehabilitation, in: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, IEEE, 2010, pp. 3690–3693.

[15] M. Khademi, H. Mousavi Hondori, A. McKenzie, L. Dodakian, C. V. Lopes, S. C. Cramer, Free-hand interaction with leap motion controller for stroke rehabilitation, in: Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems, ACM, 2014, pp. 1663–1668.

[16] K. Vamsikrishna, D. P. Dogra, M. S. Desarkar, Computer-vision-assisted palm

rehabilitation with supervised learning, IEEE Transactions on Biomedical Engineering 63 (5) (2016) 991–1001.

[17] C. R. Pereira, D. R. Pereira, F. A. Silva, J. P. Masieiro, S. A. Weber, C. Hook, J. P. Papa, A new computer vision-based approach to aid the diagnosis of parkinson's disease, Computer Methods and Programs in Biomedicine 136 (2016) 79–88.

[18] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[19] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[20] H. Chen, Q. Wang, L. Cao, Design of the workstation for hand rehabilitation based on data glove, in: Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on, IEEE, 2010, pp. 769–771.

[21] H. Yamaura, K. Matsushita, R. Kato, H. Yokoi, Development of hand rehabilitation system for paralysis patient–universal design using wire-driven mechanism–, in: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, IEEE, 2009, pp. 7122–7125.

[22] S. Amsuss, P. M. Goebel, N. Jiang, B. Graimann, L. Paredes, D. Farina, Self-correcting pattern recognition system of surface emg signals for upper limb prosthesis control, IEEE Transactions on Biomedical Engineering 61 (4) (2014) 1167–1176.

[23] M. Schultz, J. Gill, S. Zubairi, R. Huber, F. Gordin, Bacterial contamination of computer keyboards in a teaching hospital, Infection Control & Hospital Epidemiology 24 (04) (2003) 302–303.

[24] A. V. Dowling, O. Barzilay, Y. Lombrozo, A. Wolf, An adaptive home-use robotic rehabilitation system for the upper body, IEEE journal of translational engineering in health and medicine 2 (2014) 1–10.

[25] C. D. Murray, S. Pettifer, T. Howard, E. L. Patchick, F. Caillette, J. Kulkarni, C. Bamford, The treatment of phantom limb pain using immersive virtual reality: three case studies, Disability and rehabilitation 29 (18) (2007) 1465–1469.

[26] C. Schönauer, T. Pintaric, H. Kaufmann, S. Jansen-Kosterink, M. Vollenbroek-Hutten, Chronic pain rehabilitation with a serious game using multimodal input, in: Virtual Rehabilitation (ICVR), 2011 International Conference on, IEEE, 2011, pp. 1–8.

[27] D. S. Alexiadis, P. Kelly, P. Daras, N. E. O'Connor, T. Boubekeur, M. B. Moussa, Evaluating a dancer's performance using kinect-based skeleton tracking, in: Proceedings of the 19th ACM international conference on Multimedia, ACM, 2011, pp. 659–662.

[28] M. Raptis, D. Kirovski, H. Hoppe, Real-time classification of dance gestures from skeleton animation, in: Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation, ACM, 2011, pp. 147–156.

[29] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, M. Zhou, Sign language recognition and translation with kinect, in: IEEE Conf. on AFGR, 2013.

[30] L. Pigou, S. Dieleman, P.-J. Kindermans, B. Schrauwen, Sign language recognition using convolutional neural networks, in: Workshop at the European Conference on Computer Vision, Springer, 2014, pp. 572–578.

[31] O. Lopes, M. Reyes, S. Escalera, J. Gonzàlez, Spherical blurred shape model for 3-d object and pose recognition: Quantitative analysis and hci applications in smart environments, IEEE Transactions on Cybernetics (2014) 1–1.

[32] J. Uijlings, I. Duta, E. Sangineto, N. Sebe, Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off, International Journal of Multimedia Information Retrieval 4 (1) (2015) 33–44.

[33] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, International journal of computer vision 103 (1) (2013) 60–79.

[34] H. Wang, D. Oneata, J. Verbeek, C. Schmid, A robust and efficient video representation for action recognition, International Journal of Computer Vision 119 (3) (2016) 219–238.

[35] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[36] L. Liu, L. Shao, F. Zheng, X. Li, Realistic action recognition via sparsely-constructed gaussian processes, Pattern Recognition 47 (12) (2014) 3819–3827.

[37] L. Shao, X. Zhen, D. Tao, X. Li, Spatio-temporal laplacian pyramid coding for action recognition, IEEE Transactions on Cybernetics 44 (6) (2014) 817–827.

[38] D. Wu, L. Shao, Silhouette analysis-based action recognition via exploiting human poses, IEEE Transactions on Circuits and Systems for Video Technology 23 (2) (2013) 236–243.

[39] B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzalez, F. X. Roca, A selective spatio-temporal interest point detector for human action recognition in complex scenes, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1776–1783.

[40] I. Laptev, On space-time interest points, International journal of computer vision 64 (2-3) (2005) 107–123.

[41] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: European conference on computer vision, Springer, 2008, pp. 650–663.

[42] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association, 2008, pp. 275–1.

[43] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: BMVC 2009-British Machine Vision Conference, BMVA Press, 2009, pp. 124–1.

[44] C. Wu, J. Zhang, S. Savarese, A. Saxena, Watch-n-patch: Unsupervised understanding of actions and relations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4362–4370.

[45] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 588–595.

[46] D. Wu, L. Shao, Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 724–731.

[47] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, A. Erçil, A decision forest based feature selection framework for action recognition from rgb-depth cameras, in: International Conference Image Analysis and Recognition, Springer, 2013, pp. 648–657.

[48] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, L. Sigal, A. Argyros, C. Sminchisescu, et al., Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary, in: Proceedings of the 15th ACM on International conference on multimodal interaction, ACM, 2013, pp. 365–368.

[49] D. Gong, G. Medioni, X. Zhao, Structured time series analysis for human action segmentation and recognition, IEEE transactions on pattern analysis and machine intelligence 36 (7) (2014) 1414–1427.

[50] J. Luo, W. Wang, H. Qi, Group sparsity and geometry constrained dictionary learning for action recognition from depth maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1809–1816.

[51] F. Lv, R. Nevatia, Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, Computer Vision–ECCV 2006 (2006) 359–372.

[52] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[53] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence 35 (1) (2013) 221–231.

[54] L. Ge, H. Liang, J. Yuan, D. Thalmann, Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3593–3601.

[55] M. Oberweger, P. Wohlhart, V. Lepetit, Hands deep in deep learning for hand pose estimation, arXiv preprint arXiv:1502.06807.

[56] J. Tompson, M. Stein, Y. Lecun, K. Perlin, Real-time continuous pose recovery of human hands using convolutional networks, ACM Transactions on Graphics (ToG) 33 (5) (2014) 169.

[57] G. Chéron, I. Laptev, C. Schmid, P-cnn: Pose-based cnn features for action recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 3218–3226.

[58] A. Bulat, G. Tzimiropoulos, Human pose estimation via convolutional part heatmap regression, in: European Conference on Computer Vision, Springer, 2016, pp. 717–732.

[59] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, J.-M. Odobez, Deep dynamic neural networks for multimodal gesture segmentation and recognition, IEEE transactions on pattern analysis and machine intelligence 38 (8) (2016) 1583–1597.

[60] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 3361–3368.

[61] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Spatio-temporal convolutional sparse auto-encoder for sequence classification., in: BMVC, 2012, pp. 1–12.

[62] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[63] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: International Workshop on Human Behavior Understanding, Springer, 2011, pp. 29–39.

[64] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, F. X. Roca, Deep pain: Exploiting long short-term memory networks for facial expression classification, IEEE Transactions on Cybernetics.

[65] X. Yang, Y. Tian, Effective 3d action recognition using eigenjoints, Journal of Visual Communication and Image Representation 25 (1) (2014) 2–11.

[66] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele, Deepcut: Joint subset partition and labeling for multi person pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[67] M. Sun, P. Kohli, J. Shotton, Conditional regression forests for human pose estimation, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 3394–3401.

[68] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 9–14.

[69] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2752–2759.

[70] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on systems, man, and cybernetics 9 (1) (1979) 62–66.

[71] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.

[72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[73] F. Perronnin, Y. Liu, J. Sánchez, H. Poirier, Large-scale image retrieval with compressed fisher vectors, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3384–3391.

[74] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European conference on computer vision, Springer, 2010, pp. 143–156.

[75] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Computer Vision and Pattern Recognition (CVPR), 2015.
URL http://arxiv.org/abs/1409.4842

[76] G. Ozbulak, Y. Aytar, H. K. Ekenel, How transferable are cnn-based features for age and gender classification?, in: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE, 2016, pp. 1–6.

[77] O. Koller, H. Ney, R. Bowden, Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled, in: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 3793–3802.

[78] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093.

[79] P. J. Werbos, Backpropagation through time: what it does and how to do it, Proceedings of the IEEE 78 (10) (1990) 1550–1560.

[80] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, F. X. Roca, Deep pain: Exploiting long short-term memory networks for facial expression classification, IEEE Transactions on Cybernetics.

[81] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.