Quantifying Progression of Multiple Sclerosis via Classification of Depth Videos

Peter Kontschieder[1], Jonas F. Dorn[2], Cecily Morrison[1], Robert Corish[1], Darko Zikic[1], Abigail Sellen[1], Marcus DSouza[3], Christian P. Kamm[4], Jessica Burggraaff[5], Prejaas Tewarie[5], Thomas Vogel[2], Michela Azzarito[2], Ben Glocker[1], Peter Chin[6], Frank Dahlke[2], Chris Polman[5], Ludwig Kappos[3], Bernard Uitdehaag[5], and Antonio Criminisi[1]

[1]Microsoft Research (GB), [2]Novartis Pharma (CH), [3]University Hospital Basel (CH), [4]University Hospital Bern (CH), [5]VU University Medical Center Amsterdam (NL), [6]Novartis Pharmaceuticals East Hanover (US)

Abstract This paper presents new learning-based techniques for measuring disease progression in Multiple Sclerosis (MS) patients. Our system aims to augment conventional neurological examinations by adding quantitative evidence of disease progression. An off-the-shelf depth camera is used to image the patient at the examination, during which he/she is asked to perform carefully selected movements. Our algorithms then automatically analyze the videos, assessing the quality of each movement and classifying them as healthy or non-healthy. Our contribution is three-fold: We i) introduce ensembles of randomized SVM classifiers and compare them with decision forests on the task of depth video classification; ii) demonstrate automatic selection of discriminative landmarks in the depth videos, showing their clinical relevance; iii) validate our classification algorithms quantitatively on a new dataset of 1041 videos of both MS patients and healthy volunteers. We achieve average Dice scores well in excess of the 80% mark, confirming the validity of our approach in practical applications. Our results suggest that this technique could be fruitful for depth-camera supported clinical assessments for a range of conditions.

1 Introduction

Multiple Sclerosis is a chronic, inflammatory and degenerative disease of the central nervous system that affects over 2.5 million people worldwide and leads to impairment and disability over time. Treatment focuses on anti-inflammatory drugs, preventing relapses, and to a lesser extent reducing progression. The availability of a measurable progression marker is important *e.g.* to assess the effect of various drugs on a given patient. However, the current gold standard measure, the Expanded Disability Status Scale (EDSS) [9] has high inter- and intra-rater variability making change hard to quantify [8,11]. While alternatives have been proposed, none have been sufficiently validated for the use as primary outcomes.

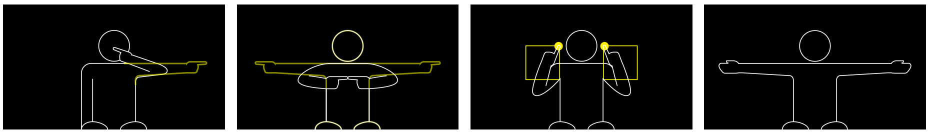

Figure 1. Illustration of the four movements used in our system. Finger-to-Nose (FNT), Finger-to-Finger (FFT), Drawing Squares (DRS) and Truncal Ataxia (TAT). See Section 3 for details.

Here, we address this problem via a depth camera system in conjunction with machine learning algorithms to quantify changes in movement-related symptoms in an objective manner. Common movement-related symptoms are *e.g.* tremor (*i.e.* a rhythmic oscillation of a body part), ataxia (*i.e.* swaying of the torso when staying up right), or lack of accuracy when trying to touch an object (*e.g.* the nose). In our system we capture videos of patients while performing carefully chosen movements (see Fig. 1). Then our algorithms analyze each video and classify them as healthy or not. Here we focus on the patients motor skills alone, *i.e.* additional impairments such as cognitive ones are not considered.

Related literature. Much work has been done on the automatic analysis of brain images for MS patients [1]. For example, the work in [7] uses random forests [3] for the automatic segmentation of brain lesions in multi-channel, MR scans. In [6] prior knowledge about the brain anatomy is integrated within a statistical framework for the classification of healthy brain tissues as well as the detection of lesions in MR images.

In this paper we take a substantially different approach, measuring the impact of MS on the patients motor abilities by imaging their movements rather than their brain. While initial attempts have been made in this area for MS [12], they relied solely on the skeleton produced by depth camera APIs rather than direct analysis of the images. To our knowledge, our work is the first example of depth video analysis for clinical assessment in MS or other movement-disabling conditions. We believe that it may represent a better assessment of how the disease affects the daily life activities of those who live with it.

At a procedural level, the patients are asked to perform expert-selected, motion-rich tasks (see Fig. 1). An off-the-shelf Kinect depth sensor records such movements into depth videos. Colour images are discarded to respect patient privacy. For the video classification step, we test and extend three ensemble-based, discriminative classifiers. One based on decision forests [3], and two based on new, randomized SVMs [14]. Resulting Dice scores (on previously unseen videos) in excess of the 80% mark confirm the validity of our approach.

2 Ensemble Learning for Depth Video Classification

This section describes how to learn ensembles of classifiers for the binary classification of depth videos into patients and healthy subjects. The challenges are the need for: i) good generalization despite training on only few training videos, ii)

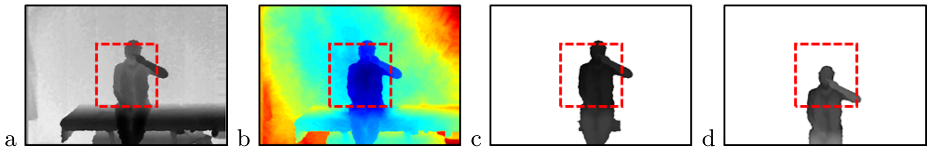

Figure 2. Preprocessing of depth video (image centre as red box). **a)** Inpainted depth image. **b)** Foreground (FG) model assignment and corresponding geodesic distances [5] (blue corresponds to small distance, here to FG class). **c)** Obtained FG segmentation. **d)** Spatial registration and depth normalization after head detection.

automatic selection of clinically relevant, discriminative spatio-temporal landmarks in the depth video, and iii) coping with variable video length (video duration may itself be a discriminative feature, *e.g.* patients may be slower).

To this end, we propose to use new variants of decision forests [3] and support vector machines (SVMs) [14] for *structured input space exploration*. With 'structured input' we refer to the fact that we have automatically registered videos. This allows our algorithms to rely on a common reference where location, depth and motion features for pixels can be compared with one another.

Depth Video Preprocessing. The video preprocessing stage performs foreground segmentation and registration to assure that the recorded persons are roughly centred in the image (see Fig. 2d). We start with inpainting depth values in regions where Kinect does not provide measurements, using nearest neighbours. Then, the closer subject is separated from the background using a Gaussian model of depths followed by a geodesic-based refinement stage [5] (see Fig. 2b). Finally, template-matching in depth space is used as a head detector for centring and mapping the segmentations to a canonical depth.

2.1 Structured Video Exploration with Classifier Ensembles

Given the segmented depth videos we need to encode them for the purpose of classification. In the related field of action recognition videos are often described via histograms of optical flow, space-time features or bag-of-features (see *e.g.* [10]). However, the most commonly used descriptors consider only local spatio-temporal intervals, making them unsuitable to capture the possibly slow anomalies that may occur in MS. Additionally, differences in how a patient or a healthy subject perform the same movement may be more subtle than differences between categories such as running, jumping and clapping, which are typical for action recognition tasks. Indeed, early stage MS patients may show very mild motion anomalies, which may still be used as evidence for early disease detection.

Visual features. Here we derive visual features to capture effects such as tremor in limbs and other motion-related instabilities, based on optical flow for consecutive image pairs in depth videos. More formally, for a segmented depth video with size $(\mathsf{w} \times \mathsf{h} \times \mathsf{d})$ we assume to be given pairs of optical flow $(V_{\mathsf{x}}(x, y, t), V_{\mathsf{y}}(x, y, t)), t \in \{1, \ldots, \mathsf{d} - 1\}$, *i.e.* flow components (x, y) for pixel positions (x, y) in two images taken at time t and $t + 1$.

To go beyond optical flow information of only image pairs and discover longer and more informative time intervals in the videos we consider the physical quantity of acceleration, describing the rate at which the velocity of an object changes with time. Hence, we are writing $A_j^T(x,y) = \mathit{\Delta} V_j(x,y)/\mathit{\Delta} T$ for acceleration component $j \in \{\mathsf{x},\mathsf{y}\}$ at position (x,y) and time interval $T = [\mathsf{d}\tau_1, \mathsf{d}\tau_2)$, $\tau_1, \tau_2 \in [0,\ldots,1)\,, \tau_1 < \tau_2$ with duration $\mathit{\Delta} T = \mathsf{d}(\tau_2 - \tau_1)$. Finally, we use the following approximation to determine the acceleration quantities in a more robust way

$$\tilde{A}_j^T(\cdot) = \sum_{t\in T} \mathbb{1}\left[\operatorname{sgn}\left(V_j(\cdot,t)\right) \neq \operatorname{sgn}\left(V_j(\cdot,t+1)\right)\right] g([V_{\mathsf{x}}(\cdot,t)\ V_{\mathsf{y}}(\cdot,t)]^\top, \kappa)\,, \qquad (1)$$

where $\mathbb{1}[P]$ is the indicator function based on P and $\operatorname{sgn}(P)$ returns the sign of P. Moreover, $g(\boldsymbol{v},\kappa) = G(\|\boldsymbol{v}\|_2 - \kappa)$ where $G(P)$ is the Heaviside step function response for P and $\kappa \in \mathbb{R}^+$ is a hyper-parameter to control the influence of the flow magnitude, *i.e.* it allows us to eliminate noisy oscillations from the flow components. In this sense, Eq. (1) provides us with a location-dependent quantity measuring sign changes in the optical flow vectors over time, which can be efficiently computed. Next, we describe how to learn the most informative spatio-temporal parameters in our structured input space classifiers.

Depth Video Exploration via Ensemble Learning. Structured video classification ensembles are classifiers with specific knowledge and constraints about the input space they are applied to. In our case, the input space $\mathcal{X} = \{(V_{\mathsf{x}}(\cdot,t), V_{\mathsf{y}}(\cdot,t)) : t \in \{1,\ldots,\mathsf{d}-1\}\}$ contains optical flow information for all consecutive pairs of preprocessed and spatially registered videos. Our output space $\mathcal{Y} = \{\mathtt{PAT}, \mathtt{HS}\}$ is binary, *i.e.* we classify a sample as patient or healthy subject. Given the standardized input space, we treat the entire video as a single sample $x \in \mathcal{X}$, *i.e.* our goal is to explore and identify the most informative spatial and temporal areas of the input space on a global level rather than learning only a local model. Also, training and testing is extremely fast, *i.e.* inference has constant complexity per classifier, independently of the video length. Next, we describe how the relevant parameters can be learned in ensembles of decision trees or support vector machines (SVMs), respectively.

Ensembles of decision trees. Random forests [3,4] are ensembles of binary decision trees and our approach can be seen as a modification over standard classification trees. More formally, the training set $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$ is split into two subsets for the left $\mathcal{Z}_{\mathtt{L}}$ and right $\mathcal{Z}_{\mathtt{R}}$ child nodes (such that $\mathcal{Z}_{\mathtt{L}} \cap \mathcal{Z}_{\mathtt{R}} = \emptyset$ and $\mathcal{Z}_{\mathtt{L}} \cup \mathcal{Z}_{\mathtt{R}} = \mathcal{Z}$) until a stopping criterion is met. Information gain Q is often used to measure the quality of parameters Θ for the binary split $\psi(x|\Theta) \to \{\mathtt{L},\mathtt{R}\}$, where $(x,y) \in \mathcal{Z}$ is an entire training video x with corresponding binary ground truth label y. Consequently, the optimal parameters can be found as $\Theta^* = \arg\max_\Theta Q(\mathcal{Z},\Theta)$ where $Q(\mathcal{Z},\Theta) = H(\mathcal{Z}) - \sum_{i\in\{\mathtt{L},\mathtt{R}\}} \frac{|\mathcal{Z}_i|}{|\mathcal{Z}|} H(\mathcal{Z}_i)$ with $\mathcal{Z}_i = \{(x,y) \in \mathcal{Z} : \psi(x|\Theta) = i\}$ and $H(\cdot)$ is the entropy estimated from the empirical probability distributions over $\mathcal{Y}$ in the resulting child nodes. From here, we populate Θ to contain the relevant parameters, *i.e.* we learn the optimal parametrizations for our visual features in (1) in terms of spatial locations and time intervals in the videos. Hence, we define $\Theta = (B_1, B_2, \gamma, d_1, d_2, k)$ where $B =$

$(x_1, y_1, x_2, y_2, \tau'_1, \tau'_2)$ defines a cuboid selecting a space-time window in the image domain where $\tau'_1, \tau'_2 \in [0, \ldots, 1)$ defines a temporal sequence on the unit interval. In addition we learn a threshold parameter $\gamma \in \mathbb{R}$, a variable $d \in \{\mathsf{x}, \mathsf{y}\}$ selecting one of the acceleration components and a variable $k \in \{1, \ldots, 4\}$ selecting one of the following functions to obtain the binary split $\psi(x|\Theta) = \mathbb{1}[\rho(x|\Theta) > \gamma]$ using

$$\rho(x|\Theta) = \begin{cases} f(d_1, B_1) & k = 1 \\ f(d_1, B_1) + f(d_2, B_2) & k = 2 \\ f(d_1, B_1) - f(d_2, B_2) & k = 3 \\ |f(d_1, B_1) - f(d_2, B_2)| & k = 4, \quad \text{and} \end{cases} \tag{2}$$

$$f(d, B) = \frac{1}{(y_2 - y_1)(x_2 - x_1)} \sum_{x'=x_1}^{x_2} \sum_{y'=y_1}^{y_2} \tilde{A}_d^{T(B)}(x', y')\,, \tag{3}$$

where $T(B) = [\tau'_1, \tau'_2]$ and $\tilde{A}(\cdot)$ stems from Eq. (1). Finally, the best node split parameters Θ^* are obtained by evaluating randomly chosen ones for Θ for a pre-defined number of trials and stored in each internal node.

Ensembles of SVMs. Support Vector Machines (SVMs) [14] are non-probabilistic, binary classifiers defined by separating hyperplanes in the feature space. In standard SVM training, the input data would be linearized versions of our input samples from $\mathcal{X}$ as defined before. However, due to the high and variable number of dimensions (*e.g.* each sample lives in a high dimensional $\boldsymbol{x} \in \mathbb{R}^{2\mathsf{wh}(\mathsf{d}-1)}$ space), training becomes very slow and requires temporal normalization which is undesired since it alters the movement characteristics. Instead, we propose a randomized input space exploration with ensembles of SVMs by adapting the above described technique for decision trees. For each SVM in our ensemble we consider a fixed size input space $\mathcal{X}^{\mathtt{SVM}}$ of much smaller dimensionality than $\mathcal{X}$. Then, each component of a sample $\boldsymbol{x}^{\mathtt{SVM}} \in \mathcal{X}^{\mathtt{SVM}}$ is populated using Eq. (2) with randomly sampled parameters for Θ as for tree node training. Consequently, the degree of video space to be explored can be controlled by the dimensionality chosen for the input data space $\mathcal{X}^{\mathtt{SVM}}$. The final predictions for both, the decision tree ensemble and the SVM ensemble is obtained by output averaging.

3 Experimental Evaluation

Here we provide a movement description and present our experimental findings.

Movements and Recording Protocol. We analyze depth video recordings for patients at different stages of their disease (considering their cerebellar function scores for upper extremities provided from expert neurologists) and healthy subjects. Still images of the movements are shown in Fig. 1 and both examined groups were advised in the same manner on how to perform them. For **Finger-to-Nose (FNT)**, the person has to stretch out the left or right arm horizontally and then move the index finger towards the nose until contact. For **Finger-to-Finger (FFT)** the person has to stretch out both arms and index fingers and bring the latter together in parallel until they touch. Both, FNT and FFT are repeated three times in total. For **Drawing Squares (DRS)** the index fingers are lifted to eye level, brought down to breast height, moved outwards toward the shoulders, up to eye level and finally back to the starting position without repetitions. For **Truncal Ataxia (TAT)** both arms are stretched out to the side and held for 5 seconds. All movements are repeated with eyes open and closed. While FNT and TAT are part of standard neurological examinations, FFT and DRS were chosen by our MS expert clinicians in order to stimulate various cerebellar functions. Over several months, we collected a total of 1041 (317 PAT, 724 HS) depth videos in two hospitals and in a separate location only for healthy subjects, using the above recording protocol.

Quantitative Analysis. We could not use [13] as a baseline since an initial data analysis phase revealed that the variance of body joint location accuracy exceeds the amplitude of the movement-related motions we intend to measure. For our proposed ensemble classifiers we used the following setup in a 5-fold cross validation. We trained 300 trees in each fold until a minimum of 2 samples were left in the leaves, used 2000 trials for node parameter estimation and an inverse frequency reweighing to correct the imbalance of available training samples for the PAT and HS classes, in our own implementation. Likewise, we trained 300 Linear SVMs and 300 Kernel SVMs using 500-dimensional input samples populated as described above and hyper-parameter optimization was done using a grid search over a 10-fold cross-validation set. We used the SVM implementation in OpenCV [2] and the optical flow of [15] therein and fixed $\kappa = 1.15$ in Eq. (1).

Movement Classification Results. For comparing our proposed classifier ensembles we present class-specific dice scores ($D_{\texttt{PAT}}$, $D_{\texttt{HS}}$) and their mean $\overline{D}$ in Tab. 1. The gray-shaded cells highlight the best performing methods for which we additionally list sensitivity $S_{\texttt{SENS}}$, specificity $S_{\texttt{SPEC}}$, the percentage of correct predictions $S_{\texttt{GLOB}}$ and corresponding standard deviations over the 5-fold cross validation for all of the above. We also report the average numbers of training/test samples per fold. For a small subset of videos (6 per movement) we made frame- and pixel-wise annotations for the raw depth videos based on which we report segmentation accuracy scores ($S_{\texttt{SENS}}$ and $S_{\texttt{SPEC}}$) after spatial registration to assess the quality of the preprocessing stage. All proposed classifier ensembles show encouraging results with FNT, FFT and DRS above 80% mean Dice score for SVMs and TAT at 74% for forests, where we hope to improve with more training data in future. The preprocessing pipeline yields almost perfect results on the videos we have randomly selected for ground truth annotation.

Automated Landmark Selection. In Fig. 3 we show discriminative landmarks learned by our forest-based algorithm as heat maps over corresponding movement images (acceleration component x on top and y on bottom, both summed over the temporal dimension). Clearly, each movement has its specific, informative regions which are in high correspondence to where swaying of the waist/torso (DRS) or shoulders/arms and head tremor (TAT) is exhibited. For DRS also the 'corners' of the square are emphasized. Moreover, our features seem very informative in the nose region for FNT or where the fingers should meet for FFT, *i.e.* areas where intention tremor is predominant. The rightmost plots in

Method	**FNT**			**FFT**			**DRS**			**TAT**		
	$\overline{D}$	$D_{\texttt{PAT}}$	$D_{\texttt{HS}}$	$\overline{D}$	$D_{\texttt{PAT}}$	$D_{\texttt{HS}}$	$\overline{D}$	$D_{\texttt{PAT}}$	$D_{\texttt{HS}}$	$\overline{D}$	$D_{\texttt{PAT}}$	$D_{\texttt{HS}}$
Forest	84.3	79.4	89.2	74.9	58.1	91.7	81.1	70.3	92.1	74.3	57.8	90.9
	±4.4	±5.9	±2.9	±6.3	±11.0	±1.7	±3.8	±5.1	±2.6	±5.7	±8.8	±2.6
Linear SVM	80.9	75.3	86.4	79.1	65.6	92.7	84.5	75.1	93.9	73.4	56.5	90.3
	±3.0	±5.0	±1.1	±4.4	±7.6	±1.2	±2.8	±4.0	±1.5	±3.7	±5.1	±2.3
Kernel SVM	85.2	80.5	89.9	81.0	68.3	93.7	81.4	70.6	92.2	66.2	45.8	86.7
	±4.0	±5.7	±2.3	±3.6	±6.6	±0.6	±2.6	±3.4	±1.8	±4.0	±7.2	±0.8
$S_{\texttt{SENS}}$ $S_{\texttt{SPEC}}$ $S_{\texttt{GLOB}}$	78.3	91.4	86.7	79.3	91.1	89.5	89.8	90.3	90.2	74.3	86.8	85.1
	±10.1	±2.2	±3.2	±17.4	±2.9	±0.9	±7.1	±3.0	±2.3	±11.9	±3.5	±3.9
Avg. #Train/fold `PAT`, `HS`	103.2	186.4		52.8	96.6		48.0	87.0		49.6	89.4	
Avg. #Test/fold `PAT`, `HS`	25.8	46.4		13.2	76.6		12.0	61.6		12.4	78.4	
Segmentation $S_{\texttt{SENS}}$ $S_{\texttt{SPEC}}$	99.9	97.9		98.2	99.9		99.9	99.8		99.8	99.8	

Table 1. Quantitative results for all experiments, all D, S in %. See text.

Fig. 3 show the importance of x/y acceleration components as a function of the unit time interval for videos FNT and FFT. There are roughly three modes for FFT with decreasing amplitude, possibly due to the repetitions of moves while for FNT the sampling is more uniform. Finally we remark that all landmark information is discovered with weak supervision only (only a single, binary label per video is available), making the decision process transparent to the clinician.

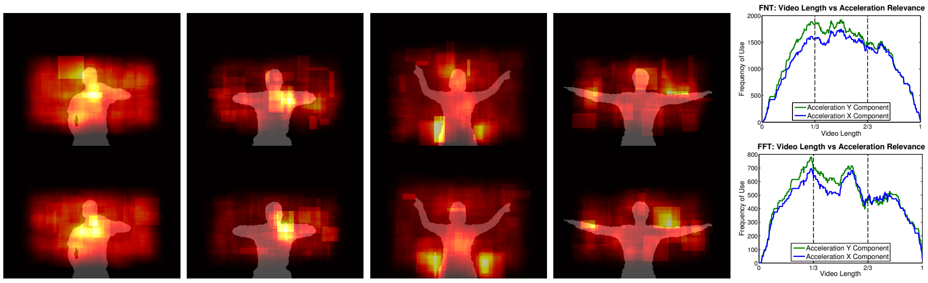

Figure 3. Automatic selection of clinically relevant landmarks. Left: Heat map visualizations of discriminative image regions for acceleration features (x on top, y on bottom) for FNT, FFT, DRS and TAT (left to right). Right: Importance of acceleration components as function of video length for FNT (top) and FFT (bottom). Please use digital zoom for better visibility.

4 Conclusions and Future Work

We have introduced a new system for quantitative assessment of disease progression for Multiple Sclerosis (MS) patients, based on a depth camera setup and a novel depth video classification approach. Using depth video recordings of neurologically relevant movements we have proposed structured input classifier ensembles to distinguish patients from healthy subjects. The idea of our classifiers is to automatically infer discriminative spatio-temporal regions within depth videos. We introduced ensembles of decision trees and SVMs to learn novel acceleration features and their parametrizations as part of the training process and evaluated them on a new dataset of 1041 depth videos from MS patients and healthy subjects. Our experimental evaluation showed encouraging performance and confirmed the fact that automated MS assessment from depth videos is possible. In the future we will investigate how to learn and fuse predictions across multiple movements and provide predictions in an ordinal continuous domain.

References

1. MICCAI Workshop on Medical Image Analysis on Multiple Sclerosis: validation and methodological issues (MIAMS) (2009)
2. Bradski, G.: Opencv. Dr. Dobb's Journal of Software Tools (2000)
3. Breiman, L.: Random forests. In: Machine Learning. vol. 45, pp. 5–32 (2001)
4. Criminisi, A., Shotton, J.: Decision Forests in Computer Vision and Medical Image Analysis. Springer (2013)
5. Criminisi, A., Sharp, T., Blake, A.: Geos: Geodesic image segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) Computer Vision – ECCV 2008, Lecture Notes in Computer Science, vol. 5302, pp. 99–112. Springer Berlin Heidelberg (2008)
6. Datta, S., Narayana, P.A.: A comprehensive approach to the segmentation of multichannel three-dimensional mr brain images in multiple sclerosis. (PAMI) 2 (2013)
7. Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N.: Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance. Neuroimage (2011)
8. Goodkin, D., Cookfair, D., Wende, K., Bourdette, D., Pullicino, P., Scherokman, B., Whitham, R.: Inter- and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke expanded disability status scale (EDSS). Neurology 42(4), 859–859 (1992)
9. Kurtzke, J.F.: Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology 33(11) (1983)
10. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: (CVPR) (2008)
11. Noseworthy, J.H., Vandervoort, M.K., Wong, C.J., Ebers, G.C.: Interrater variability with the expanded disability status scale (EDSS) and functional systems (FS) in a multiple sclerosis clinical trial. Neurology 40(6) (1990)
12. Pfueller, C., Otte, K., Mansow-Model, S., Paul, F., Brandt, A.: Kinect-based analysis of posture, gait and coordination in multiple sclerosis patients. Neurology 80 (2013)
13. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A.: Efficient human pose estimation from single depth images. (PAMI) (2013)
14. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
15. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: Proc. DAGM Symposium (DAGM). pp. 214–223 (2007)