Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision

Advait Sarkar[1,2], Cecily Morrison[2], Jonas F. Dorn[3], Rishi Bedi[3], Saskia Steinheimer[4], Jacques Boisvert[3], Jessica Burggraaff[6], Marcus D'Souza[5], Peter Kontschieder[4], Samuel Rota Bulò[4], Lorcan Walsh[3], Christian P. Kamm[4], Yordan Zaykov[2], Abigail Sellen[2], Siân Lindley[2]

[1]University of Cambridge
Cambridge, United Kingdom
advait.sarkar@cl.cam.ac.uk

[2]Microsoft Research
Cambridge, UK
cecilym@microsoft.com

[3]Novartis Pharma AG
Basel, Switzerland
jonas.dorn@novartis.com

[4]Department of Neurology
University Hospital Bern
Bern, Switzerland

[5]Department of Medicine
University Hospital Basel
Basel, Switzerland

[6]VU University Medical Center
Amsterdam, Netherlands

ABSTRACT

A growing number of domains, including affect recognition and movement analysis, require a single, real number ground truth label capturing some property of a video clip. We term this the provision of continuum labels. Unfortunately, there is often an uncacceptable trade-off between label consistency and the efficiency of the labelling process with current tools. We present a novel interaction technique, 'setwise' comparison, which leverages the intrinsic human capability for consistent relative judgements and the TrueSkill algorithm to solve this problem. We describe SorTable, a system demonstrating this technique. We conducted a real-world study where clinicians labelled videos of patients with multiple sclerosis for the ASSESS MS computer vision system. In assessing the efficiency-consistency trade-off of setwise versus pairwise comparison, we demonstrated that not only is setwise comparison more efficient, but it also elicits more consistent labels. We further consider how our findings relate to the interactive machine learning literature.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI)

Author Keywords

Setwise Comparison; Continuum Labels; Video; Computer Vision; Machine Learning; Interactive Machine Learning; Health

CHI'16, May 07–12, 2016, San Jose, CA, USA.

ACM 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858199

INTRODUCTION

Many computer vision applications are built to produce a real number output based on an input video. To train these applications requires a single, real number ground truth label per video that captures the property to be predicted. For example, the magnitude of positive valence in a video of a smile for affect recognition applications [31]; the clinical impression of 'tremor' in a video of a patient performing a standard movement to train a system for tracking neurological disease progression [23]; or the degree of correctness of a rowing stroke in automated sports coaching [15]. In contrast to other video labelling situations, these cases require only a single judgement, but that judgement necessitates seeing continuous change in movement across the entire video. We refer to these types of labels as continuum labels.

We address continuum labels which differ in three ways from labelling challenges already tackled in computer vision research (e.g. security surveillance): (1) many domains require expert labellers, making crowdsourcing in its current form untenable; (2) the labels are continuous (e.g., real numbers in the interval $[0, 1]$), rather than discrete categories (e.g., 0, 1, 2), making label consistency a challenge; (3) the property of interest can only be perceived temporally, i.e., needs to be observed over time in each video, which makes rapid (efficient) comparisons difficult. As reliable labels are critical to the performance of the resultant machine-learned model, we need better approaches to support labelling in application domains reliant on video training data with continuum labels.

Providing continuum labels is tedious. Large databases, in the order of hundreds of examples, or better thousands, are needed to serve as ground truth training sets. There is substantial work in the computer vision community on creating tools for labelling, such as the FEELTRACE system [12]. However, unlike tools that may be developed in the field of human-computer interaction, these tools do not specifically build on human capabilities. As a result, potential interaction solutions

to problems perceived as intractable in the computer vision community, such as continuum labels, are less explored.

The continuum labelling problem is well exemplified by challenges faced in the development of ASSESS MS, a computer vision system for the assessment of motor ability in patients with multiple sclerosis [33]. It aims to provide a more consistent and fine-grained measure of motor ability to enable reliable tracking of disease progression, since neurologists exhibit high inter- and intra-rater variability in their assessments. Ultimately, this high variability which motivated ASSESS MS also necessitated the development of a more consistent way to capture clinical judgement to be used as ground truth labels.

There is a trade-off between efficiency and consistency when doing continuum labelling. For example, providing discrete numeric labels to categorize motor ability, e.g. 0 to 4, is efficient, requiring only a few seconds per label, but inter- and intra-labeller reliability are consistently low even with training [11]. One way of improving consistency is to use pairwise comparison, for which a person is asked whether one entity is better, worse, or the same as another. After comparing all possible pairs, a ranked order can be calculated. In earlier experiments, we have found that this increases consistency, but is inefficient. The number of pairwise comparisons required grows quadratically with the number of videos, and so does not scale to the number of videos needed for the kinds of computer vision applications on which we focus here.

To address the problem of scalable, consistent continuum labelling, we introduce setwise comparison, a novel interaction technique that makes consistent labelling tractable at the scale required for computer vision applications. Like pairwise comparison, setwise comparison builds on the cognitive ability of labellers to provide better relative judgements than absolute ones, but achieves better efficiency by asking labellers to make sets of relative judgements, from which a complete ranking can be inferred using Bayesian techniques, specifically the TrueSkill algorithm. The interaction modes employed draw upon interactive machine learning techniques, extending them for continuum labels.

This paper makes the following contributions:

- We introduce the problem of continuum labelling through preliminary studies conducted while developing the ASSESS MS system.
- We present SorTable, a system which implements setwise comparison, a novel interaction technique that makes consistent continuum labelling tractable.
- We describe a real-world comparative study of pairwise and setwise comparison on ASSESS MS patient data, demonstrating that setwise comparison produces significantly more consistent and efficient labels.
- We discuss how our findings may be applied to other interactive machine learning applications.

THE ASSESS MS LABELLING PROBLEM

Motivation: Improving Standard Clinical Assessment

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system causing a wide range of symptoms including ataxia (swaying of the body), tremor (oscillation of a body part), numbness, paralysis, visual disturbance, and various forms of cognitive difficulties. Symptoms, either alone or in combination, are usually suffered during relapses followed by extended periods of remission in which symptoms may improve. Over time the disease can enter into a progressive phase in which a steady deterioration occurs, affecting the ability to do everyday tasks, such as walking or eating [22]. The unpredictability of the disease course makes the ability to track MS particularly useful.

The condition is currently assessed using the Expanded Disability Status Scale (EDSS) [26], a standard rating scale based on clinical examination. Patients are asked to perform a range of functional exercises, including stretching out one arm to the side and then touching the nose (Finger Nose Test) or walking on a pretend tight rope (Tightrope Walking). Patients are then observed for specific symptoms, such as tremor, which the clinician rates on an integer scale that usually ranges from 0 to 4. Multiple such sub-scores are summarized into a total EDSS score from 0 to 10. Although the EDSS is widely used and accepted, it suffers from low inter- and intra-rater reliability making disease tracking problematic [34]. Moreover, the coarse-grained nature of the sub-scores, meant to increase rater agreement, makes the scale less sensitive to changes in patient state.

ASSESS MS aims to address this problem by quantifying changes in motor ability more consistently and with finer granularity than currently possible. A Kinect camera captures depth+colour videos of neurological assessment movements performed by patients in a clinical setting with the support of the health professional. These videos are pre-processed to isolate the patient and movement is captured through an alphabet of motion filters learned in an unsupervised manner. Features derived through this alphabet of filters are then utilised in a supervised machine learning approach coupled with EDSS sub-score labels.

Problem: Defining Ground Truth

ASSESS MS requires reliable ground truth labels if it is to provide a consistent measure of motor ability. In an early attempt to create consistent ground truth labels, we developed a protocol in which all clinical team members $(n = 4)$ jointly agreed on labels for a set of 100 "gold-standard" videos based on the relevant EDSS sub-score. Specifically, clinicians blind-labelled videos and then explicitly verbalized their label criteria. This process was expected to lead to a joint, stable concept for labelling. Each clinician then labelled several hundred videos that included a sample of "gold standard" videos from their own clinic. Labelling was relatively efficient (median rating time of 3.3s per video). Consistency, however, was poor, ranging from 23-69% agreement on the previously agreed "gold standard" videos.

To increase label consistency, we explored the method of pairwise comparison. Clinicians were shown pairs of patient videos and asked to choose which exhibited higher disability or whether they were equal. When 11 neurologists performed pairwise rating of 50 videos, an interclass correlation coefficient (ICC) of 0.71 was achieved, which if averaged across 2 raters, increased to a 0.96 median. In the medical literature, this is considered a highly robust level of consistency [19]. This method also provided a finer level of granularity than the existing EDSS scale, with clinicians distinguishing at least one additional level of motor ability within each existing integer sub-score.

While pairwise comparison solved the problem of higher consistency and finer-grained assessment, it failed in terms of efficiency. Comparing every video to every other in the set of 50 required 1,225 comparisons. Neurologists required between 87 and 146 minutes to rate this set of videos. As the number of comparisons grows quadratically with the number of videos, this is clearly not tractable for even a small video set of 300, which would require 44,850 comparisons, or over 37h of continuous labelling. The need for a method for which the labelling effort grows linearly with the number of videos drove the development of the setwise comparison technique.

RELATED WORK

Continuum Labelling

Labelling is a key part of supervised machine learning algorithms [24]. Large training datasets are needed in which each training data point has a label, enabling statistical judgement of what a data point outside of the training set might be labelled. By continuum labelling we refer to the provision of labels that consist of a single, continuous value that captures a certain property of a video, which may summarise other continuously changing properties across all frames in that video. This could be a numeric score or probability and is typically associated with regression problems. It stands in contrast to categorical labels, typically associated with classification problems.

Scale is a key challenge for labelling. This is a particular problem for video clips, which are typically more time-consuming to perceive and compare, and therefore to label, than static media (e.g. images). While companies often pay people to label data (e.g. for search engines), crowdsourcing is being explored as a more cost-effective alternative. A number of tools have been built to specifically support efficient frame-by-frame annotation of video in which people and objects are identified and tracked [38]. There are specific tools to enable crowd-support in coding behavioural actions in video [28]. The use of "surrogates" (static summaries) to capture key elements of a video has also been explored [32]; and attempts have been made to automatically label realistic human actions in movies using natural language script [27].

Unfortunately, none of these approaches is appropriate to continuum labelling in the domains that we have identified, which require expert knowledge, such as that of athletic coaches [15] or doctors [33]. Even applications that draw on common knowledge, such as affect recognition, often require experts to achieve consistent continuum labels [3]. People may be able to easily distinguish between boredom and laughter, but determining the difference between some laughter and a lot of laughter in a consistent manner takes expertise. As a result, small numbers of people must label large numbers of samples, often at high cost, as existing techniques to increase speed, such as surrogates, are not applicable given the continuous nature of what is being labelled.

Another substantial problem in the literature is label noise, or labels that do not accurately reflect the data point [16]. Many approaches have been developed to address this problem in the machine learning literature. For instance, gaining repeated labels [21], modelling labeller noise [39], or automatic detection of outlier labels [9]. The first two approaches are not appropriate to small numbers of expert labellers. The final one assumes mislabelled data points can be statistically characterised. However, mislabelling in continua often arises not because the label is inherently wrong, but because the concept being labelled is "fuzzy", that is, without clear boundaries.

The support that people need in evolving a concept in order to label it consistently has been identified and addressed for static data in the interactive machine learning literature [25]. We review this literature below.

Interactive Machine Learning

The interactive machine learning literature focuses on how best to integrate expert human judgement into machine learning systems. To achieve this, users visit and revisit a sequence of examples on which to provide human judgement [13, 17]. These systems amplify the abilities of people to make judgements that are useful to a machine learning system. They do this by supporting a number of design objectives [5]: (1) helping the user pick the next example to judge; (2) helping the user to decide what judgement to give; and (3) helping the user see how their judgements impact the machine learning model. We are interested in the second objective, despite a stronger focus in the literature on the first and third.

Concept evolution suggests that the boundaries between labels can shift depending on context. For instance, in a website topic classification task, what exactly constitutes a website about travel, and how does it differ from a website for expatriates? The lack of clarity of where the boundaries are can lead to inconsistent labelling of training data which decreases the accuracy of machine learning results. Techniques such as structured labelling [25] help people evolve and concretize concepts through repeated viewing and categorization of examples with reference to other examples, supporting boundary definition.

The systems that support concept evolution reported in the literature draw upon visual salience in static media to help people organize their concepts and in doing so scaffold the decision making process. For example, structured labelling [25] allows users to spatially organize groups and tag them while determining whether they are part of a particular concept. CueFlik [14] enables people to form image search queries by bringing together training examples in the same spatial frame. BrainCel [35] provides multiple coordinated views of the model being built to support machine learning in spreadsheets. Capitalis-

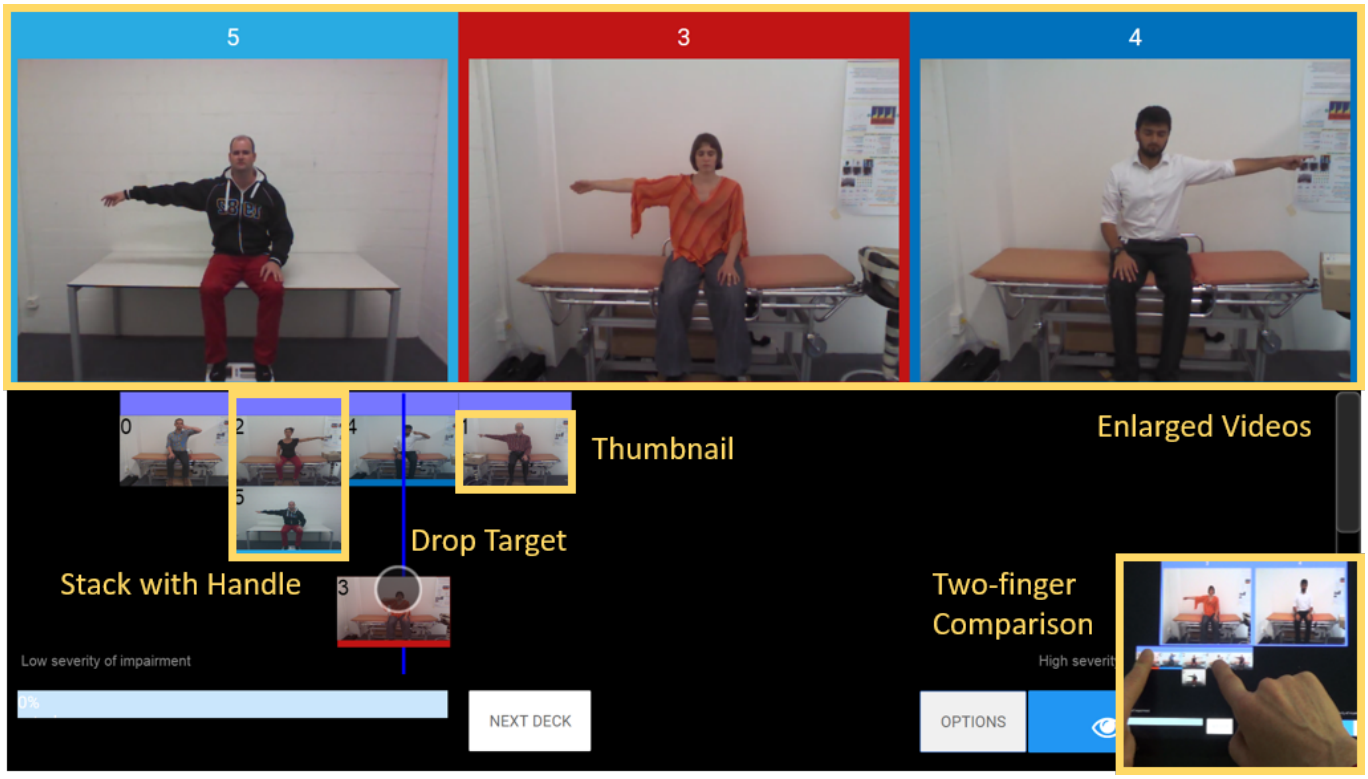

Figure 1. SorTable system.

ing on visual salience in static media is potentially useful for continuum labelling, but we needed to extend it to video and address the issue of scale to apply it to our problem.

Paired Comparison

Pairwise comparison, described earlier, is a technique to support users in their decision-making through relative preference judgements. People are given paired data points and asked to express preference or ranking, such as "better than." After every combination of pairs has been labelled, a ranked order of preference can be determined. This technique, also referred to as two-alternative forced choice, has been used substantially in the psychology literature, common approaches including Bradley [8] and Thurstone [36]. There have subsequently been many variations and applications of this approach, from understanding the extent of visual perception [2] to understanding decision-making [7]. Pairwise comparison is closely linked in psychological literature with another set of techniques generally grouped together under the name of card-sorting. These techniques ask participants to physically group cards as a method of concept articulation. One example is *Q sorting*, which enables the systematic study of subjectivity, such as a person's viewpoint, opinion, or beliefs [37].

Pairwise comparison has also been used as a mechanism for determining labels. Carterette et al [10] demonstrate that this method can facilitate judgements beyond binary ones for information retrieval applications, but the method does not scale to large datasets even when optimisations are made to the pair selection strategy. Preference judgements have also been used as the basis for a game to motivate labelling of data [6]. In this case, label agreement is achieved between cooperative players by reducing the number of images shown until agreement is reached.

SYSTEM DEVELOPMENT

SorTable System Design

SorTable is our interface for labelling videos to be used in machine-learning applications that rely on expert-provided continuum labels. It was designed to improve the efficiency-consistency trade-off by enabling setwise comparisons of videos. Users are given multiple sets of videos to sort on a virtual touch-enabled table, ordering them from least to greatest or stacking them to indicate equality. This design builds upon the human ability to make consistent relative judgements (as demonstrated in pairwise comparison) and is inspired by the use of visual salience in concept evolution tools.

Sets are not independent of each other. Every set shares a proportion of its constituent videos with at least some other sets. This enables us to apply the TrueSkill algorithm [20], originally developed for ranking gamers, which aggregates the sorted sets into a rank order of all videos. It does so through a probabilistic Bayesian approach, which efficiently predicts pairwise judgements between videos which have never co-occurred in a set.

SorTable aims to increase efficiency without losing substantial consistency through three specific features. First, the presentation of videos in sets builds upon human short-term memory to make multiple comparisons at once. Second, the ability

to create stacks to indicate that videos are the same can substantially reduce the number of comparisons as the labeller need only refer to one video in the stack when sorting. Third, SorTable facilitates mixed-strategy sorting, including the automatic display of the left and right neighbours of the currently selected video, and the ability to compare any two videos with a two-finger gesture. All interactions are touch based.

System Description

Sets of videos are presented to the user as thumbnails, with three enlarged for viewing as in Figure 1. The center video is highlighted in red and its left and right neighbours in light and dark blue respectively. Tapping any thumbnail selects that video as the central one, displaying its respective neighbours. Thumbnails can be dragged for sorting, snapping to a regular grid when dropped. A 'drop target' line indicates the position they will assume when the finger is lifted. The enlarged videos themselves can dragged left and right for reordering.

Videos can be stacked by dragging thumbnails on top of, or underneath, one another to indicate that they are equal. There is no limit to stack size. Entire stacks can be dragged using the stack handles above each stack. Any two videos can be compared on an ad-hoc basis with a two-finger gesture. With the left finger touching one thumbnail, the right finger can be used to tap on other thumbnails comparing several videos with the first in rapid succession. The standard behaviour of showing neighbours can be resumed by tapping anywhere else on the screen. Thumbnails that have not been touched are marked with a bright blue circle in the corner. This fades continuously over three touches, helping ensure that labellers do not leave videos uninspected.

We used the multiplayer TrueSkill algorithm variant with ties, which treats the equality judgements arising from stacks with mathematical rigour [20, 1]. The TrueSkill algorithm was chosen over alternatives such as the Bradley-Terry [8] or Thurstone models [36], more commonly used in psychology literature, as it is well suited for situations with few raters (clinicians) and many items to rate (patients). It has faster convergence, naturally handles draws (no difference between patients) and enables more than two comparisons at a time (setwise). Each set was considered an 8-player free-for-all game. Following standard practice for applying TrueSkill, each video's "score" was initialised with an arbitrary default mean of 25 and standard deviation of 25/3. These scores are internal to TrueSkill, and are unrelated to the standard clinical EDSS sub-score our system ultimately produces.

The system is implemented using web technologies. Any set of videos can be loaded, along with a specification file containing additional parameters such as set size. We used a Lenovo Yoga touchscreen laptop for development, and conducted our studies on a $27''$ touchscreen with 1920×1080 HD resolution.

Key Design Decisions

To explore key design decisions, we undertook initial usability testing with three of our ASSESS MS team neurologists. Three sessions were held over a 4-week period. Interaction with the system was observed and audio recorded, ensuring that no patient data was captured in observance of hospital rules.

We also ran a number of simulations to explore potential parameter values. Using a dataset of 50 videos in which we had previously established a reliable ground truth through pairwise comparison done by 11 neurologists, we simulated set outcomes using a subset of the pairwise comparison data as a representation of a single neurologist's judgment.

This initial work helped us understand: (1) ideal composition of the set, including number and type of videos; and (2) the usability of the comparison interactions we provided.

Set Composition

During usability testing, we tried sets of 5 as well as 10 patient videos. Neurologists felt that sets of 5 were too easy, especially if there were many with the same label, while sets of 10 were highly taxing. The neurologists felt sets of 7 to be ideal.

Set selection is also influenced by the set overlap size, that is, the number of videos per set that have already appeared in another set. A larger overlap size provides more effective pairwise comparisons and should be more accurate (as more information is shared between sets, which aids TrueSkill's estimation), but it also increases the total number of sets to be labelled. We performed a parameter search using simulations to select set size, overlap size, and set composition strategy. Figure 2a demonstrates that the Pearson correlation between the known ground truth and the simulated games changes as a function of set size and overlap size.

We also explored how the similarity of videos in terms of disease severity that composed a set would affect both human and algorithm sorting performance. We considered specifically whether we should provide sets with similar or varied initial ratings previously provided by a clinician. Similar sets could potentially increase speed by enabling more stacks and more efficient use of TrueSkill. Varied sets, however, were less cognitively taxing for clinicians during usability testing and completed more quickly. Figure 2 shows that more varied sets result in higher correlation to ground truth.

Balancing the clinicians' reported experience and the results of the simulations, we chose to have a set size of 8, an overlap of 3, and a varied strategy for set composition.

Comparison Interactions

There are three comparison interactions. The user can choose one video for display with its immediate neighbours, encouraging constant pairwise comparison between neighbouring videos. We considered enabling participants to independently choose three videos for comparison, but felt this would unnecessarily increase interactional friction. Our approach of enforcing which videos were available to compare employs a design-with-intent philosophy [29]. Usability testing showed that this supported a strategy in which the user started at one 'end' of the set and watched all videos in a systematic order through to the other 'end' before sorting.

We developed a second comparison interaction for more ad-hoc video comparison after our first usability test. We noted that one participant would compare a video against several others to determine where it should be placed or stacked. Without an explicit way to choose a second video, the participant

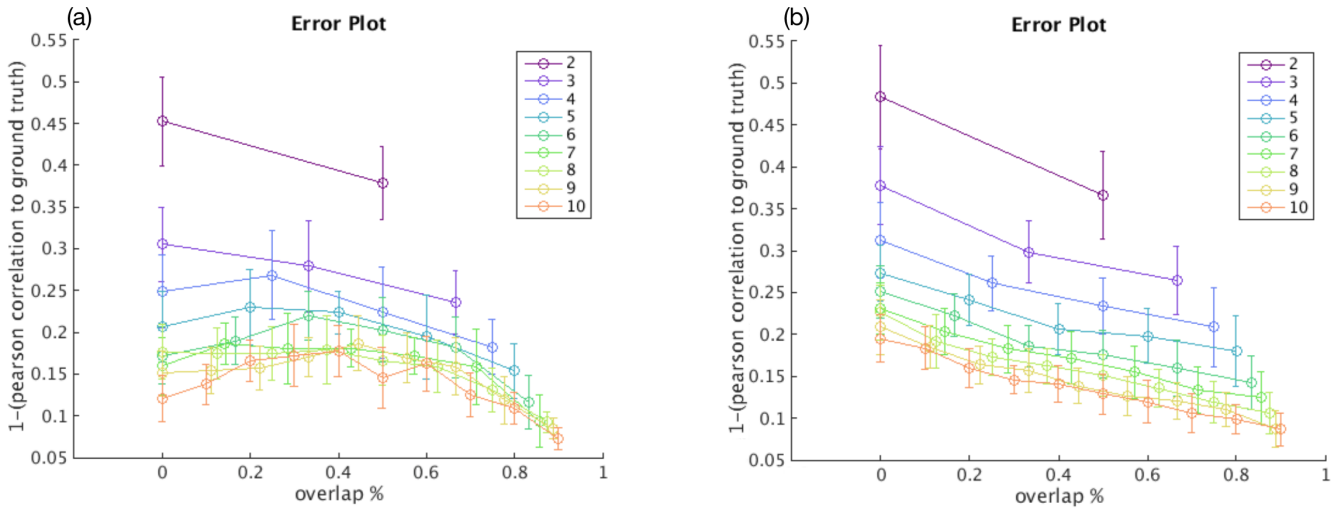

Figure 2. Illustration of the correlation of the simulated games (y-axis) to the known ground truth for different set size (graphed lines) and overlap size (x-axis): (a) videos with similar disease severity; (b) videos with varied disease severity.

had to rely on their memory or temporarily make two videos neighbours. We therefore added the two-finger gesture to enable rapid comparison of two specific videos, allowing us to accommodate this less-frequent, but nonetheless important, alternative flow.

The third comparison approach enabled the use of stacks to demonstrate equality of judgement. In our initial usability testing, two of three clinicians found stacks useful. Stacks are also efficient as they reduce the effective set size and thereby cognitive load. To encourage the use of stacks, we implemented a 'stack handle' following the first usability session. This allows users to move complete stacks around with the handle located at the top of stack.

EVALUATION

Study Design

We conducted a within-subjects study to compare the efficiency-consistency trade-off between pairwise comparison and setwise comparison. More specifically, we assess whether we improve the efficiency of the labelling task with an acceptable deterioration in label consistency. We asked the following research questions:

- RQ1: Does setwise comparison improve the efficiency-consistency trade-off when compared to pairwise comparison?
- RQ2: Is the cognitive load of setwise comparison comparable or less than pairwise comparison?
- RQ3: What interaction strategies were employed during setwise comparison?

Recruitment

A convenience sample of eight neurologists with experience with multiple sclerosis and not involved in the ASSESS MS project team were recruited. They covered a wide span of experience levels (Median: 6y, Range: 1-26y) and age (Median: 35y, Range: 28-54y). None had prior experience labelling patient videos. The participants were all from the same hospital as it is not permitted to take the patient videos offsite.

Protocol

Participants were asked to rate a set of 40 patient videos using both pairwise and setwise comparison. The SorTable system was used in both cases, with the set size specified at 2 and 8 for the pairwise and setwise conditions respectively. Videos of the Finger Nose test, a neurological test of tremor and upper body dysmetria, were used. The same videos were compared by all participants in both conditions to enable exact calculation of consistency and efficiency in a small sample size. To reduce the possibility of clinicians remembering patients, we had a minimum of three days between conditions. Starting condition was counterbalanced.

Before each condition, a short pre-scripted tutorial was read out by the study facilitator and the participant was prompted to try each interaction technique. Before beginning the study, two practice sets were done to reduce learning effects. Following the completion of the comparisons, the participants were asked to fill out a questionnaire that included the full NASA Task Load Index [18] as well as three open-ended questions.

Analysis

To address RQ1, we calculated consistency and efficiency for each condition. Consistency was evaluated first through a global intraclass correlation coefficient (ICC type "A-1" [30]), which reports the absolute agreement of scores across different labellers on a scale from 0 (no agreement) to 1 (perfect agreement). We also calculated average ICC, by averaging the Pearson correlation between every combination of two labellers for each condition. Efficiency was measured as the total time the participant needed to complete the task. Average ICCs were compared by an unpaired t-test, while mean time to completion was tested for equality by a paired t-test. Normality was established with the Shapiro-Wilk test.

	Global ICC	Average ICC	Task Time (min)
		mean±sd [min−max]	mean±sd [min−max]
Pairwise	0.70	$0.77 \pm 0.1 [0.64 - 0.94]$	$77.86 \pm 14.53 [52.79 - 95.84]$
Setwise	0.83	$0.85 \pm 0.07 [0.72 - 0.95]$	$23.80 \pm 8.12 [13.28 - 36.06]$
t-test		$p = 5 \cdot 10^{-4}$	$p = 4 \cdot 10^{-5}$

Table 1. Consistency and efficiency of pairwise and setwise comparison of 40 videos.

We used the full NASA Task Load Index to measure cognitive load in both conditions to address RQ2. It is a validated, multi-dimensional scale which provides a self-reported measure of overall workload associated with a task. Participants rate six dimensions on an analog scale with 20 unnumbered sections: mental demand, physical demand, temporal demand, performance, effort, and frustration. A weighted average of these is then calculated after participants rank the perceived importance of each dimension in creating cognitive load through pairwise comparison. We compare both the overall score of cognitive load as well as each dimension separately (i.e. the raw TLX calculation), using paired t-tests. In addition to a formal measure of cognitive load, we asked the participants three open-ended questions: (1) What three words would you use to describe this system to a colleague? (2) What did you like most about the system? (3) What did you like least about the system? This data was synthesized and reported in summary form.

Addressing RQ3, we look at the frequency of the three special interaction techniques: video swiping (to change order); stacking, and two-finger comparison. We then looked holistically at the sorting process, visualizing how the thumbnails in each set were re-ordered over time. We manually inspected these visualizations for interaction patterns.

Results

Efficiency-Consistency Trade-off

Setwise comparison was significantly more efficient than pairwise comparison. Task time was reduced by an average of 54 minutes ($p = 4 \cdot 10^{-5}$). Although a set took a median of 127.8s and 4.2s for setwise and pairwise comparison respectively, the smaller number of sets in setwise comparison (10 vs 858) made a substantial difference to total task time for all participants as shown in Figure 3. Setwise comparison also had a lower total task time standard deviation than pairwise comparison, 6.9m versus 14.5m, suggesting higher predictability of task time.

As shown in Table 1, setwise comparison is also more consistent than pairwise comparison. Setwise comparison has both a higher global intraclass correlation coefficient (ICC) as well as a higher average ICC ($p = 5 \cdot 10^{-4}$) among pairs of labellers. Consistency of pairwise labelling is similar to values in our earlier experiments. Our data shows no trend of improved consistency in the second condition, suggesting that there was no meaningful learning effect between the two conditions.

Figure 3. Total task times.

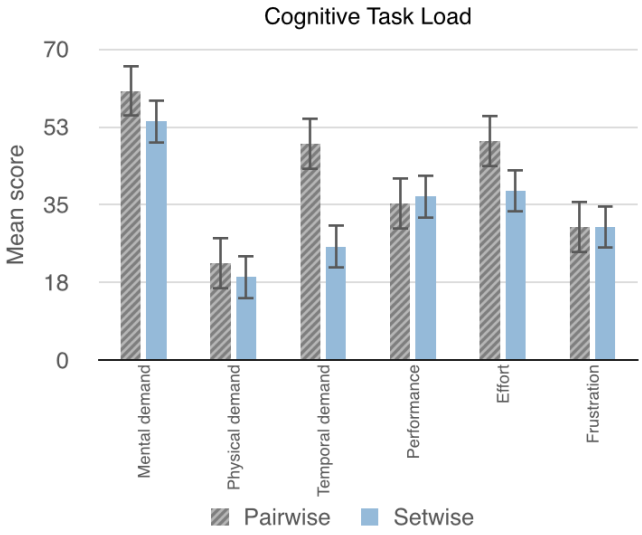

Figure 4. Mean scores for raw TLX measures.

Cognitive Load

There were no significant differences between the setwise and pairwise conditions on the NASA Cognitive Task Load Index or on any of the six dimensions that it comprises. We see surprisingly similar scores as shown in Figure 4, except for a notable reduction in temporal demand and effort for setwise comparison.

Responses to the open-ended questions were positive, without being particularly revealing. Participants mentioned that the

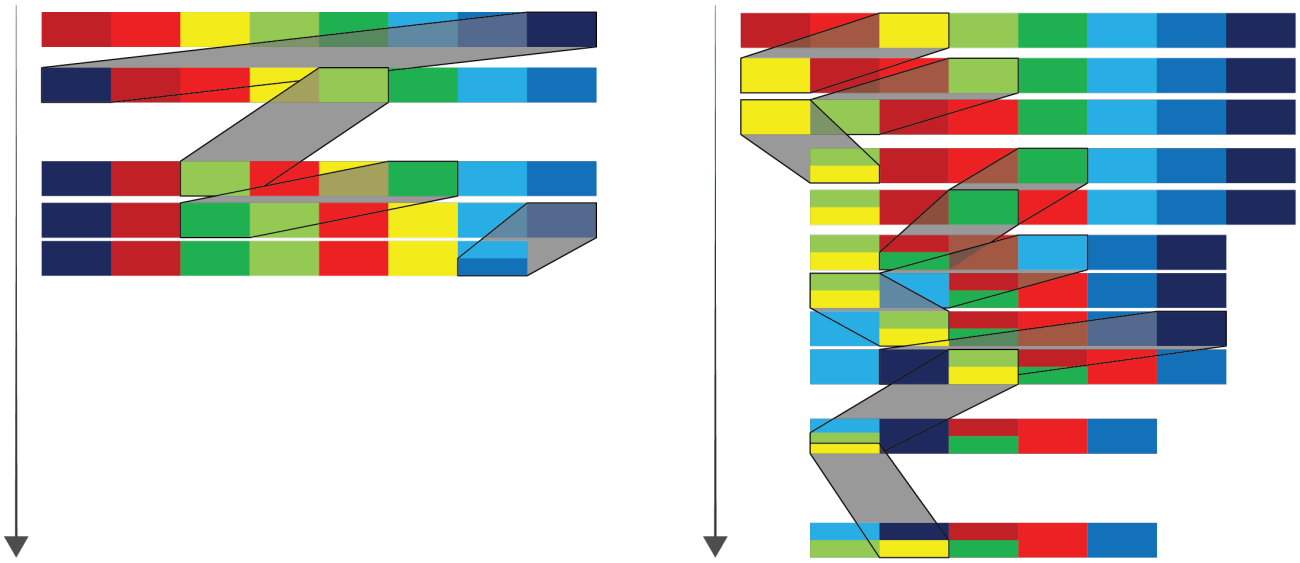

Figure 5. Contrasting sorting strategies employed during setwise comparison. Left: "insertion"-style sorting; Right: "bubble"-style sorting. Each colour represents a thumbnail and the y-axis represents time. Stacks are depicted by shared cell.

software was intuitive and easy to handle. Most participants commented on the use of a touchscreen device and the ability to drag and drop efficiently. There was a common complaint that videos were too small, with proposition that the screen real estate could have been used differently or that an option to zoom could have been available. The comments made were the same for pairwise and setwise comparison.

Interaction Strategies

We proposed three interaction strategies in our design. We found that swapping of neighbouring enlarged videos was done by 3 labellers in 60% of their sets. Swaps were most commonly done in the beginning of the labelling process, but one person used it throughout. The two-finger comparison gesture was used by 5 of the labellers, 4 regularly. The gesture was used an average of 3.2 times per set with a range of 0-12 times. Stacks were also heavily used with only 6 (of 80) sets not containing stacks. Most sets (53%) had only one stack per set. There were, however, 11 instances in which a set had three stacks. Most commonly stacks contained two videos, but ranged up to six videos.

Through visualizing the activity, we consider whether there are distinguishable strategies for comparing videos in sets. We see a marked difference in the number of interactions, which included moving videos using thumbnails, moving stacks, swapping the enlarged videos in the upper interface, and engaging the two-finger comparison gesture. On one extreme, participant 7 had only 153 interactions over 10 sets, while on the other extreme, participant 3 had 323 interactions, over twice that of participant 7. The other participants ranged in between and we could group them as follows. Low interaction: participants 2 and 7 (median moves: 14.5-15); Moderate interaction: participants 4, 5, 6, 8 (median moves: 20-23.5); and High interaction: participants 1 and 3 (median moves: 29-32).

The interaction levels reflect distinct strategies that can be seen visually in Figure 5. Some participants viewed a large number of videos and relied on their memory to then sort them, interacting with most videos only once. This was often achieved by dragging videos that were clearly on the edge of the spectrum before ordering the middle videos, a strategy which is similar to the insertion sort algorithm. Others did rapid comparisons and swaps of immediate neighbours, bubbling videos to the correct position, reminiscent of the bubblesort algorithm.

The neurologists were fairly consistent in their strategy, but strategies did not have a clear connection to experience level or starting condition. There was also no clear connection between strategy and completion time or consistency. For example, participants 2 and 3 had equal completion times but did 171 versus 323 actions, respectively. Likewise, participants 1 and 7 had similarly high consistency (.88 and .9 respectively), but were in the high and low interaction categories respectively.

DISCUSSION

Setwise Comparison

We introduced the concept of setwise comparison to make the relative judgements that provide consistency in pairwise comparison tractable at the scale needed for continuum labelling in computer vision. Setwise comparison was significantly more efficient than pairwise comparison. If we consider how these two methods scale from 40 videos to 400 videos, a modest training set size for these kinds of applications, pairwise comparison would take approximately 100 hours, whereas setwise comparison would take only $\sim$3.5 hours. The substantial improvement in efficiency comes not only in faster completion times, but in the use of the TrueSkill algorithm so that the method scales linearly. We would expect this method to scale up to thousands of videos, a reasonable training data set for these types of applications.

As is evident from our research questions, our initial assumption was that there would be a trade-off between consistency and efficiency between setwise and pairwise comparison. We only expected to approximate the consistency of labels yielded by pairwise comparison. We were surprised that setwise comparison was not only more efficient, but also significantly more consistent. The setwise ICC of 0.83 is considered excellent in the medical field, defined as a result above 0.8 [19]. While pairwise comparison improves when labels are combined across labellers, setwise comparison stays relatively steady with high average ICC across labellers. This suggests that multiple labellers are not needed, a particular advantage for a task requiring expert labellers.

We also saw no difference in cognitive load either globally or on specific dimensions, suggesting that setwise comparison was not substantially more mentally demanding than pairwise comparison. We had expected that the cognitive demand on short-term memory might make setwise comparison harder, but this was not the case. The fact that mental demand is not highlighted in this test inspires confidence that consistency may scale well to large data sets.

We have presented positive results in a comparison of pairwise and setwise approaches, albeit with a small sample. It is possible that we might have had more pronounced differences on the cognitive load measures with a larger sample size. Nonetheless, the size is realistic to continuum labelling tasks which require expert labellers. We would expect that our system could be applied to other continuum labelling problems with a similar set-up. It may be that data requiring less detailed judgement may allow for larger sets or an increased number of enlarged videos.

The fact that our data cannot be taken out of the hospital in which it is generated indicates next steps. For example, it would be interesting to explore how we might combine data ranked in different settings (which cannot be shared) into a complete ranking to be used in ASSESS MS. This raises further research directions, such as the development of an appropriate overlap strategy that could enable the combination of data sets using TrueSkill.

Reflections for Interactive Machine Learning

The development of the SorTable system for setwise comparison builds on the interactive machine learning literature, especially concept evolution. A key aspect of concept evolution is to address that concepts being used as ground truth to a machine learning algorithm may shift with further comparison. This is manifested in our scenario in that the boundaries of a particular rating category (e.g., the boundary between a motor ability label of '1' versus '2') can never be exactly determined. SorTable dispenses with the need for people to have such delineated concepts of a motor ability, by requiring only a relative relationship between videos. While the problem we solve here differs from most interactive machine learning ones, which aim to produce a reusable model, there is something to be learned from reflecting on our approach.

Structured labelling, a concept evolution approach, primarily utilises visual salience of static media and the re-visiting of examples to enable consistent labels. Visual salience becomes problematic when the issue of interest is temporally displayed as in our case. Our log data suggests that concept evolution was achieved in several ways. Some participants used the specific gestures of swapping neighbours or using the comparison gesture to enable direct comparison and decision-making. Others relied on visual cues and short term memory provided by the thumbnails. Although different strategies, in both cases people were able to efficiently scaffold their own decision-making process without an impact on mental demand.

Scale for video labelling is also something that remains challenging in the interactive machine learning literature, but was necessarily designed for in setwise comparison. We considered how a person's abilities could be amplified by offloading some of the work onto the machine. We do this with the TrueSkill algorithm. Instead of having a person label every point, we ask them to sort data points and provide labels for only a few. The algorithm infers the rest of the labels. Tools have been developed for labelling action in videos that use the machine to select important key frames to reduce the scope of the task [4]. This is an equivalent notion. Concept evolution could focus more on designating relationships, such as transitivity, rather than trying to gain one label per data point.

CONCLUSION

It is increasingly common for novel applications to build upon machine learning. This necessitates the tedious task of labelling hundreds or thousands of data points to provide a ground truth dataset with which to train the model. We illustrate why this is challenging for computer vision domains such as movement or affect analysis, which require single, real number labels for videos, or what we call continuum labels. We proposed a solution, setwise comparison, on the basis that humans have greater abilities for relative judgements than absolute ones. Setwise comparison was implemented in the SorTable system, with which we demonstrated that the the setwise comparison process resulted in significantly better label consistency and efficiency than pairwise comparison for ASSESS MS patient videos.

For the most part, computer vision researchers have been responsible for finding ways to obtain labelled data for their applications. However, what we demonstrate in this paper is that through the application of good interaction design principles, and building on an understanding of human capabilities, we have been able to effectively address an important problem previously left untackled in the computer vision literature because of its apparent intractability.

ACKNOWLEDGEMENTS

We would like to acknowledge the neurologists who participated in this study: Monika Bühlmann, Verena Blatter Arifi, Julius Hodak, Nadia Di Fabio, Ariane Cavelti, Rebekka Kurmann, Andrea Seiler, Felix Riether. We'd also like to thank the ASSESS MS Steering Committee: Frank Dahlke, Antonio Criminisi, Ludwig Kappos, and Bernhard Uitdehaag.

REFERENCES

1. 2016. TrueSkill Python Code. http://trueskill.org/. (2016). Accessed: Thursday 18th February, 2016.
2. C.K. Abbey and M.P. Eckstein. 2002. Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments. *Journal of vision* 2, 1 (2002), 66–78.
3. S. Afzal and P. Robinson. 2014. Emotion Data Collection and Its Implications for Affective Computing. In *The Oxford Handbook of Affective Computing*. 359–369.
4. K. Ali, D. Hasler, and F. Fleuret. 2011. Flowboost – appearance learning from sparsely annotated video. In *IEEE computer vision and pattern recognition (CVPR)*.
5. Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney S Tan. 2011. Effective End-User Interaction with Machine Learning. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (2011), 1529–1532.
6. Paul N Bennett, David Maxwell Chickering, and Anton Mityagin. 2009. Learning consensus opinion: mining data from a labeling game. In *Proceedings of the 18th international conference on World wide web*. ACM, 121–130.
7. R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J.D. Cohen. 2006. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review* 113, 4 (2006), 700.
8. RA Bradley. 1952. Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons. *Biometrika* 39 (1952), 324–345.
9. Carla E. Brodley and Mark A. Friedl. 1999. Identifying mislabeled training data. *Journal of Artificial Intelligence Research* (1999), 131–167.
10. Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. 2008. Here or there preference judgments for relevance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4956 LNCS (2008), 16–27. DOI: http://dx.doi.org/10.1007/978-3-540-78646-7_5
11. Jeffrey A Cohen, Stephen C Reingold, Chris H Polman, Jerry S Wolinsky, International Advisory Committee on Clinical Trials in Multiple Sclerosis, and others. 2012. Disability outcome measures in multiple sclerosis clinical trials: current status and future prospects. *The Lancet Neurology* 11, 5 (2012), 467–476.
12. R. Cowie, S. Douglas-Cowie, E. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time.. In *ISCA tutorial and research workshop (ITRW) on speech and emotion.*
13. Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. *Proceedings of the 8th international conference on Intelligent user interfaces - IUI '03* (2003), 39. DOI:http://dx.doi.org/10.1145/604050.604056
14. James Fogarty, Desney S Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (2008), 29. DOI: http://dx.doi.org/10.1145/1357054.1357061
15. Simon Fothergill, Robert Harle, and Sean Holden. 2008. Modeling the model athlete: Automatic coaching of rowing technique. In *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 372–381.
16. B. Frénay and M. Verleysen. 2014. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (2014), 845–869.
17. Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, and Kevin McIntosh. 2014. You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems. *IEEE Transactions on Software Engineering* 40, 3 (2014), 307–323. DOI:http://dx.doi.org/10.1109/TSE.2013.59
18. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.
19. R. D. Hays, R. Anderson, and D. Revicki. 1993. Psychometric considerations in evaluating health-related quality of life measures. *Quality of Life Research* 2, 6 (dec 1993), 441–449. http://link.springer.com/article/10.1007/BF00422218
20. Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill(TM): A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems (NIPS2006)*. 2006.
21. P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28, 2 (2014), 402–441.
22. Christian P Kamm, Bernard MJ Uitdehaag, and Chris H Polman. 2014. Multiple sclerosis: current knowledge and future outlook. *European neurology* 72, 3-4 (2014), 132–141.
23. Peter Kontschieder, Jonas F Dorn, Cecily Morrison, Robert Corish, Darko Zikic, Abigail Sellen, Marcus D'Souza, Christian P Kamm, Jessica Burggraaff, Prejaas Tewarie, and others. 2014. Quantifying Progression of Multiple Sclerosis via Classification of Depth Videos. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 429–437.
24. S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. *Informatica 31* (2007), 249 – 268.

25. Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (2014), 3075–3084. DOI: http://dx.doi.org/10.1145/2556288.2557238

26. John F Kurtzke. 1983. Rating neurologic impairment in multiple sclerosis an expanded disability status scale (EDSS). *Neurology* 33, 11 (1983), 1444–1444.

27. Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning Realistic Human Actions from Movies. In *IEEE conference on computer vision and pattern recognition CVPR*. 1–8.

28. Walter S Lasecki, Mitchell Gordon, Steven P Dow, and Jeffrey P Bigham. 2014. Glance : Rapidly Coding Behavioral Video with the Crowd. In *Proceedings of UIST'14*. 1–11.

29. Dan Lockton, David Harrison, and Neville Stanton. 2008. Design with Intent: Persuasive Technology in a Wider Context. In *Persuasive Technology*. Springer Berlin Heidelberg, Berlin, Heidelberg, 274–278. DOI: http://dx.doi.org/10.1007/978-3-540-68504-3{_}30

30. Kenneth O McGraw and Seok P Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods* 1, 1 (1996), 30.

31. G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing* 3, 1 (Jan 2012), 5–17. DOI: http://dx.doi.org/10.1109/T-AFFC.2011.20

32. F. Metze, D. Ding, E. Younessian, and A. Hauptmann. 2013. Beyond audio and video retrieval: topic-oriented multimedia summarization. *International Journal of Multimedia Information Retrieval* 2, 2 (2013), 131–144.

33. C. Morrison, K. Huckvale, B. Corish, J. Dorn, P. Kontschieder, K. O'Hara, ASSESS MS Team, A. Criminisi, and A. Sellen. 2016. Assessing Multiple Sclerosis with Kinect: Designing Computer Vision Systems for Real-World Use. *To appear in Human-Computer Interaction* (2016). http://research.microsoft.com/apps/pubs/default.aspx?id=255951

34. JH Noseworthy, MK Vandervoort, CJ Wong, and GC Ebers. 1990. Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial. *Neurology* 40, 6 (1990), 971–971.

35. Advait Sarkar, Mateja Jamnik, Alan F. Blackwell, and Martin Spott. 2015. Interactive visual machine learning in spreadsheets. In *Visual Languages and Human-Centric Computing (VL/HCC), 2015 IEEE Symposium on*. IEEE, 159–163.

36. LL Thurstone. 1927. A law of comparative judgment. *Psychol Rev* 34 (1927), 273–286.

37. Job Van Exel and Gjalt de Graaf. 2005. Q methodology: A sneak preview. http://www.qmethodology.net/PDF/Q-methodology. (2005). Accessed: Thursday 18th February, 2016.

38. Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling. *International Journal of Computer Vision* 101, 1 (2013), 184–204. DOI: http://dx.doi.org/10.1007/s11263-012-0564-1

39. Y. Yan, R. Rosales, G. Fung, M. W. Schmidt, G. H. Valadez, L. Bogoni, L Moy, and J. G. Dy. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. (pp. 932-939).. In *International conference on artificial intelligence and statistics*. 932 – 939.