Evaluation of video activity localizations integrating quality and quantity measurements

Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al.

Evaluation of video activity localizations integrating quality and quantity measurements

Christian Wolf[1,2,a], Eric Lombardi[1,4,a], Julien Mille[1,4,a], Oya Celiktutan[1,2,5,b], Mingyuan Jiu[1,2,a], Emre Dogan[2,6a], Gonen Eren[6c], Moez Baccouche[1,2,a], Emmanuel Dellandréa[1,3,a], Charles-Edmond Bichot[1,3,a], Christophe Garcia[1,2,a], Bülent Sankur[5,b]

[1] *Université de Lyon, CNRS*
[2] *INSA-Lyon, LIRIS, UMR CNRS 5205, F-69621, France*
[3] *Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France*
[4] *Université Claude Bernard Lyon 1, LIRIS, UMR5205, F-69622, France*
[5] *Boğaziçi University, Dept. of Electrical-Electronics Engineering, Istanbul, Turkey*
[6] *Galatasaray University, Dept. of Computer Engineering, Istanbul, Turkey*

[a] {firstname.lastname}@liris.cnrs.fr
[b] {firstname.lastname}@boun.edu.tr
[c] {geren@gsu.edu.tr}

Corresponding author : christian.wolf@liris.cnrs.fr

Abstract

Evaluating the performance of computer vision algorithms is classically done by reporting classification error or accuracy, if the problem at hand is the classification of an object in an image, the recognition of an activity in a video or the categorization and labeling of the image or video. If in addition the detection of an item in an image or a video, and/or its localization are required, frequently used metrics are *Recall* and *Precision*, as well as ROC curves. These metrics give quantitative performance values which are easy to understand and to interpret even by non-experts. However, an inherent problem is the dependency of quantitative performance measures on the quality constraints that we need impose on the detection algorithm. In particular, an important quality parameter of these measures is the spatial or spatio-temporal overlap between a ground-truth item and a detected item, and this needs to be taken into account when interpreting the results.

We propose a new performance metric addressing and unifying the qualitative and quantitative aspects of the performance measures. The performance of a detection and recognition algorithm is illustrated intuitively by performance graphs which present quantitative performance values, like *Recall*, *Precision* and *F-Score*, depending on quality constraints of the detection. In order to compare the performance of different computer vision algorithms, a representative single performance measure is computed from the graphs, by integrating out all quality parameters. The evaluation method can be applied to different types of activity detection and recognition algorithms. The performance metric has been tested on several activity recognition algorithms participating in the ICPR 2012 HARL competition.

Keywords: Performance evaluation, performance metrics, activity recognition and localization, competition

1. Introduction and related work

Applications such as video surveillance, robotics, source selection, video indexing often require the recognition of actions and activities based on the motion of different actors in a video, for instance, people or vehicles. Certain applications may require assigning activities to one of the predefined classes, while others may focus on the detection of abnormal or infrequent unusual activities. This task is inherently more difficult than more traditional tasks like object recognition in static images, for a number of reasons. Activity recognition requires space-time segmentation and extraction of motion information from the video in addition to the color and texture information. Second, while object appearances in static scenes also vary under imaging conditions such as viewpoint, occlusion, illumination, the variability in the temporal component of

human actions is even greater, as camera motion, action length, subject appearance and style must also be taken into account. Finally, the characteristics of human behavior are less well understood.

Early work in this area had focused on classification of human activities, and the first works classified videos where one subject performed a single type of action. More recently, research has focused on more realistic and therefore challenging problems involving complex activities, including interactions with objects and/or containing multiple people and multiple activities. Detecting and localizing activities have therefore become as important as their classification. Evaluating detection and localization performance is inherently not straightforward and goes beyond simple measures like classification accuracy.

Indeed, evaluation of algorithms for the detection and localization of acting subject(s) within a scene is a non-trivial task. Typically, a detection result is evaluated by comparing the spatial support of the detected entity (a bounding box or a list of bounding boxes corresponding to a region in space-time) with its ground-truth space-time support. The commonly used measures, *Recall*, *Precision* and *F-Score*, must be computed in terms of the overlap proportions of these two supports. However, these measures have a serious limitation: depending on the way they are calculated, they either convey information on (i) the correctly detected proportions of the spatial support of the entity of interest, i.e., a qualitative evaluation, or (ii) the correctly detected proportion of the set of entities, i.e., a number of entities, a quantitative evaluation measure. In other words, quantitative measures relate to the recall and precision figures of activities; qualitative measures relate to how reliably activities are detected, how much of their spatial/temporal supports are recovered. It is easy to see that (ii) depends on (i), as the amount of correctly recognized entities depends on the detection quality we require for a recognition to be considered as correct. This paper addresses these issues.

The key contributions of the paper are the following:

- A new evaluation procedure is proposed for action localization which separately measures detection quality and detection quantity, and which identifies the dependency between these two concepts.

- Performance graphs are introduced that show the changes in quantity as a function of quality. The usefulness of these graphs to characterize the behavior of detection and localization algorithms is shown over recent algorithms.

Figure 1: Samples frames from one of the videos of the LIRIS / ICPR HARL 2012 dataset, as shot from a camera mounted on a mobile robot. This example contains 3 actions : 2 discussion actions (one on the blackboard, one between two sitting people), and one person typing on a keyboard. Cluttering motion is produced by other people in the background (last row).

- A single performance measure is proposed, which integrates out quality constraints and which enables the ranking of different algorithms.

- Soft upper bounds for the ranking measure and for the performance graphs are estimated from experimental data containing multiple annotations.

- Experiments show that the ranking measure is robust to annotator noise, that is variations among different annotators, while keeping a high discriminative power.

- The LIRIS human activities dataset is introduced. It has been designed specifically for the problem of recognizing complex human actions from depth data in a realistic surveillance setting and in an office environment. It has already been used for the ICPR 2012 human activities recognition and localization competition[1] (HARL). Figure 1 shows some example frames from this dataset.

- We briefly describe the entry algorithms in the ICPR 2012

[1] `http://liris.cnrs.fr/harl2012`

HARL competition and we report the evaluation results of the proposed performance metric[2] over these entries, as well as over other baseline algorithms.

The rest of this section describes existing related metrics in the literature for activity recognition and the datasets which employ them. In Section 2, our main contributions, namely, the performance metric and the performance graphs are introduced. Section 3 describes the LIRIS / ICPR 2012 HARL dataset, and section 4 illustrates the application of the proposed evaluation metric to the competition entries. Section 5 concludes.

1.1. Related metrics and datasets

Standardized performance metrics and datasets are invaluable for experimental assessment and performance comparisons of different algorithms, to guide the selection of proper solutions in practical applications. Much work has been done in an effort to generate a standard testbed for action detection and recognition systems.

Metrics — Arguably the most widely used measures for performance comparison of algorithms and datasets in the computer vision community are (i) *Accuracy*, as calculated from a confusion matrix, and (ii) *Precision*, *Recall* and the resulting *F-measure*. The former is only applicable to pure classification problems where detection and localization do not come into play. The latter measure both detection and recognition performance, and indirectly the localization performance. However they depend on certain quality constraints where a given detection must be sufficiently reliable in order to be taken into account.

A measure related to the *Precision*, *Recall* and *F-measure* class is *Receiver Operating Characteristics* (ROC) curves. These curves plot the true positive rate (related to *Recall*) versus the false alarm rate (related to *Precision*) parametrically as a function of the detection threshold. While these curves are very useful to illustrate the behavior of a method's performance over a range of operating parameters, they have two limitations. First, they can only be applied in cases where the evaluated methods can be controlled in some way, or when a confidence measure is available for each detection. Second, ROCs are applicable to binary decision problems.

Examples of cases where accuracy was used to reflect classification performance are the early datasets, such as *KTH* [1], *Weizmann* [2], *Hollywood* [3], *Hollywood-2* [4], *Olympic Sports* [5] and others. In these datasets, each video corresponds to a single action from some class, which needs to be recognized.

Criteria of the *Precision*, *Recall*, *F-measure* variety measure correct detection performance (the number of items detected) in terms of Recall, and false alarm rate (the clutter generated by imprecise detection.)

The earliest attempts for standardized performance evaluation were the *Video Analysis and Content Extraction project* (*VACE*) [6] and the *Performance Evaluation of Tracking and Surveillance* workshop series (PETS) [7]. The aim of *VACE* project was detecting and tracking text, faces and vehicles in video sequences, where two performance metrics were used [8]: a spatial frame-level measure and a spatio-temporal measure, based on the overlap between the detected object and the ground truth in the space and spatio-temporal domains, respectively. The *PETS* workshop series focused on object tracking as well as event recognition and crowd analysis. Performance metrics were defined in terms of the number of frames in which the object was tracked, the overlap between bounding boxes and the average chamfer distance. In the same vein, the *TRECVid* series [9] proposed an evaluation protocol based on temporal alignment and the two measures, called *Detection Cost Rate* (DCR) and *Detection Error Tradeoff* (DET). While DCR was defined as a linear combination of missed detections and false alarms, the temporal alignment relied on the Hungarian algorithm to find a one-to-one mapping between the system output and ground truth. The *ETISEO* project (*Evaluation du Traitement et de l'Interpretation de Sequences Video*) evaluated the results with several criteria amongst which were object localization, object shape quality, tracking time, object ID persistence and object ID confusion. The results were given in the form of ROC curves. In the *CLEAR* project [10], the metrics used in *VACE* were improved by splitting accuracy and localization error into two separate measures for a detailed failure analysis. Finally, a recent survey on the performance evaluation of vision based human activity recognition can be found in [11].

An interesting special case are action similarity based metrics, a principle introduced for the ASLAN dataset in [12]. Instead of assigning each activity to one of a (possibly large) set of classes, pairs of activities taken and classified as *same* or *not same*. This approach has several advantages: the ambiguity inherent in partitioning a set into multiple classes is addressed; the test set can contain actions which are very different from the ones in the training set; and, finally, similarity search is an

[2]The term *metric* used in the context of performance evaluation is only loosely related to the mathematical meaning of the term *metric*. In particular, the triangular inequality is not supposed to hold for metrics in this context.

application in itself, for instance in retrieval scenarios. On the other hand, the recognition of a specific activity class may be required for certain applications, as for instance surveillance and user interfaces. A classification problem can of course be solved through similarity learning, as done in [12]. However, depending on the specific task, no clear winner can be declared between direct classification and classification through similarity learning.

For completeness we also mention a class of related problems, namely detection and recognition of continuous activities. Here, the unit of evaluation is the unsegmented whole video, in which continuous streams of activities can occur. Evaluation of this variant needs metrics adapted to the problem. In [13], a measure is proposed based on alignment. It introduces six different error types: insertion, deletion, merge, fragmentation, underfill and overfill. These errors are consolidated into a segment error table, which can also be visualized in a diagram as a percentage of the total duration [14]. An improvement of this diagrams makes the performance measures invariant to class skew, as different activities by nature can have different duration times [15]. In [16], several measures are calculated on different levels: frame-level, event-level and segment-level. In [17] *Recall*, *Precision* and *F-Score* are calculated from lower level error measures like substitution, occurrence, timing and segmentation. However, these error metrics focus on temporal aspects ignoring the spatial location of activities.

All the metrics described above necessarily need somehow to integrate the detection quality measures, which are in our case determined through spatial/temporal overlap of action bounding boxes, to obtain informed quantitative measures of the number of actions detected.

Datasets — Human activity recognition in videos has a wide range of application areas such as biometrics, content-based video analysis, security and surveillance, human-computer interaction, forensics, and ambient intelligence. These different focuses have spawned several different types of datasets. The available datasets have been extensively studied in a recent survey [18]. Here, we only recall the most prominent ones.

The earliest datasets have focused on simple periodic actions, e.g., running, walking, boxing, hand-clapping etc. with usually uniform background and static camera. Each video sequence included a single person performing only one action. Typical examples are the *KTH* dataset [1] and the *Weizmann* dataset [2]. Presently, these datasets seem to be saturated in that the performances of the most recent methods reached or approach 100% accuracy. More complex actions and cluttered and dynamic backgrounds are part of the *CAVIAR* [19], *ETISEO* [20], *UIUC* [21] and *MSR* [22] action datasets, where the recordings took place in shopping centers, hallways, metro stations or in streets.

More realistic datasets include videos of a series of actions or concurrent actions performed by one or more person. These activities or events are closer to the ones in real-world scenes and are generally collected for surveillance purposes. In this context, sample datasets that focus person-person interaction are *CAVIAR* [19], *BEHAVE* [23], *CASIA* [24], *i3DPost* [25], *TV Human Interactions* [26], *UT-Interaction* [27], *VideoWeb* [28] datasets. Several datasets feature crowd behavior, for instance *PETS 2009* [29], *ETISEO* [20], or group activities, for instance *BEHAVE* [23] and *Collective Activity* [30]. Person-object interactions were addressed by *CASIA* [24], where the object can be a car, door, telephone, baggage etc. Finally, daily activities in a natural kitchen environment are dealt by the *University of Rochester Activities of Daily Living Dataset* [31] and the relatively more challenging *TUM Kitchen* [32] dataset.

Multi-view datasets include several simultaneous views for each scene: *BEHAVE* [23], *CASIA* [24], *CAVIAR* [19], *ETISEO*, [20], *IXMAS* [33], *i3DPost* [25], *MuHaVi* [34], *UCF-ARG* [35], *VideoWeb* [28] and *Multiple Cameras Fall* [36]. Aerial views are handled by *UCF Aerial* [37] and *UCF-ARG* [35].

Many datasets can be defined as "controlled" in that they are collected within the framework of a defined experimental setup. Uncontrolled databases, on the other hand, are collected without any constraints, and they are appropriately called sometimes "actions in the wild". Recently, datasets collected from Youtube, dailymotion and broadcast television channels, and movies have aroused a lot of interest. First, because they provide more realistic and challenging scenes, and second, due to the huge amount of web sources in contrast to the laborious process of building controlled databases. These datasets exhibit much larger variability as compared to the controlled datasets in their background, camera view angle, camera motion, resolution, illumination, environmental conditions, etc. and also include confounding factors such as randomness in the action rate, style, posture and clothing of the subjects. Prominent examples of wild datasets are ASLAN [12], *BEHAVE* [23], *HMDB51* [38], *Hollywood* [3], *Hollywood-2* [4], *Olympic Sports* [5], *TV Human Interaction* [26], *UCF Youtube* [39], *UCF Sports*, *UCF 50* [40] and *UCF 101* [41].

The recent introduction of low-cost depth cameras, e.g., Microsoft Kinect, Asus Xtion, Primesense Carmine and Capri,

has created wide spread interest in activity recognition from depth sequences. Depth data potentially mitigates the limitations encountered in the presence of uncontrolled lighting, camera view variations, camera motion and complex colored backgrounds etc. Processing data in 3D also makes alternative representations possible, based on depth maps or point clouds. The downside is that the current technology only allows detecting objects within a short distance to the depth sensor, i.e., it is reliable within 3-4 meters. There is a considerable amount of publicly available 3D datasets, so-called "RGB-D" or "multi-modal" datasets, in the literature. Among these one can mention *MSR Gesture 3D* [42], *MSRC-12 Kinect Gesture Dataset* [43] and the *ChaLearn Gesture Dataset* [44]. The latter focuses on gesture recognition and body sign language understanding. Basic actions such as jumping, hand clapping, stand up etc. are handled in the *Berkeley Multimodal Human Action Database (MHAD)* [45] as well as in the *Florence3D Dataset* [46]. The recognition of daily activities is addressed in the *Cornell Human Activities dataset* [47], the *RGBD-HuDaAct dataset* [48] and the *MSR Daily Activity 3D dataset* [42]. Finally, person-person interactions are provided in the *SBU-Kinect-Interaction dataset* [49].

Up to our knowledge, only two datasets exist, which contain spatial annotations in form of bounding boxes, and which are activity recognition datasets (as opposed to object tracking datasets with event detection components, like the afore mentioned PETS [7] series and others). These two datasets are the *Hollywood Localization Dataset (HLD* [50] and the *Coffee and Cigarettes Dataset (CC)* [51]. They both contain the starting and end frame number, as well as a single bounding box for a single frame of each activity. For the HLD it is the middle frame, whereas for the CC dataset it is the frame where the hand touches the head in the drinking and smoking activities. The limited spatial information is sufficient for the activities targeted by the two datasets. However, in our targeted and more complex scenarios, people move and the camera may move. In this configuration, continuous localization is important.

The LIRIS Human Activities dataset described in this paper addresses and combines several issues, providing a realistic and complex dataset featuring the following aspects and degrees of difficulty: (i) multi-modality, since both RGB and D channels are available; (ii) human-human interactions, human-object interactions and human-human-object interactions; (iii) a moving camera installed on a mobile robot; (iv) similar action classes which require integration of context; (v) full localization information with bounding boxes available for each individual frame of each activity.

2. The performance metric

We propose a new performance metric for algorithms that detect and recognize complex activities in realistic environments. The goals of these algorithms are:

- To detect relevant human behavior in midst of motion clutter originating from unrelated background activity, e.g., other people walking past the scene or other irrelevant actions;
- To recognize detected actions among the given action classes;
- To localize actions temporally and spatially;
- To be able to manage multiple actions in the scene occurring in parallel in space and in time.

The ground truth data has been annotated by marking labeled bounding boxes in each frame of each action. In particular, we assume that the ground truth annotation has segmented action occurrences, grouping all frames and bounding boxes of any one action. In other words, an action consists of a list of bounding boxes, where each bounding box corresponds to a frame. Actions consist of consecutive frames, and no frame drops are allowed in the sequence. Detection results are assumed to be in the same format. This makes it possible to provide more meaningful *Recall* and *Precision* values — indeed, a *Recall* of 90% is easier to interpret if it precisely tells us that 90% of the actions have been correctly detected.

Without this segmentation, performance measures would need to be computed on frame level and therefore would be ambiguous. In absence of segmented activities, the example of a *Recall* of 90% on frame level could be interpreted as anything of the following possibilities :

- 90% of the action bounding boxes have been correctly detected on 100% of the activities; a very unlikely case;
- 100% of the action bounding boxes have been correctly detected on 90% of the activities, a very unlikely case;
- a mixture between the first two cases; this is the general case.

The goal of the evaluation scheme is to measure a match between the annotated ground-truth and the outcome of an algorithm, i.e., between:

- A list G of ground truth actions $G^{v,a}$, where $G^{v,a}$ corresponds to the a^{th} action in the v^{th} video and where each action consists of a **set** of bounding boxes $G_b^{v,a}$ marked with one and the same class.

- A list D of detected actions $D^{v,a}$, where $D^{v,a}$ corresponds to the a^{th} action in the v^{th} video and where each action consists of a **set** of bounding boxes $D_b^{v,a}$ marked with one and the same class.

The objective is to measure the degree of similarity between the two lists. The measure should penalize two aspects, first, information loss, which occurs if whole actions or their spatial or temporal parts of actions have not been detected, and second, information clutter due to false alarms or bounding box detections which are in excess of the ground-truth. The proposed measure is inspired by a similarity measure used for object recognition in images [52], and is designed to satisfy the following goals:

1. The metric should provide a quantitative evaluation: it should indicate how many actions have been detected correctly, and how many false alarms have been created.

2. The metric should provide an indication of the quality of detection and should be easily interpretable.

The two goals, namely, to be able to determine the number of actions in the scene, and to be able to measure their detection quality are interrelated. Indeed, the number of actions we consider as detected depends on the quality threshold which we impose for any action in order to be considered as detected. A natural way to combine these two goals is first described briefly below, and then formalized in more detail in the rest of this section:

The traditional measures, *Precision* and *Recall*, quantitative measures of detection performance, form the basis of the proposed metric. In our formulation, we employ these measures with two types of threshold that gauge the amount of overlap between a ground truth action and a detected action:

1. A threshold on *pixel-level recall*, which specifies the amount of overlap between the area of detected action and the area of the ground-truth action;

2. A threshold on *pixel-level precision*, which specifies how much spurious detected area (not part of the ground truth) is allowed.

The plots of precision and recall, which depend upon the quality parameters, i.e., the thresholds, visually describe the interrelationship of quantitative and qualitative aspects of an algorithm. These are similar to the performance graphs used in [52], which relate *Recall* and *Precision* to quality thresholds.

2.1. Precision and Recall for localized activities

The first measure, *Recall*, describes the number of correctly detected action occurrences with respect to the total number of action occurrences in the dataset. The second measure, *Precision*, penalizes false alarms, by measuring the proportion of correctly detected actions within the total number of detected actions:

$$\begin{aligned} \text{Recall}(\boldsymbol{G},\boldsymbol{D}) &= \frac{\text{Number of correctly found actions}}{\text{Number of actions in the ground truth}} \\ \text{Precision}(\boldsymbol{G},\boldsymbol{D}) &= \frac{\text{Number of correctly found actions}}{\text{Number of found actions}} \end{aligned} \tag{1}$$

where G denotes ground-truths and D detections.

In order to get a single measure, these measures are combined into the traditional *F-score* [53]. The rationale of considering the harmonic mean of precision and recall is that the smaller of the two performance values is emphasized:

$$\text{F-Score}(\boldsymbol{G},\boldsymbol{D}) = \frac{2\cdot\text{Precision}(\boldsymbol{G},\boldsymbol{D})\cdot\text{Recall}(\boldsymbol{G},\boldsymbol{D})}{\text{Precision}(\boldsymbol{G},\boldsymbol{D})+\text{Recall}(\boldsymbol{G},\boldsymbol{D})} \tag{2}$$

In our modified version, these criteria involve thresholds that qualifies if and when an action can be considered as detected. Thus we can gauge how close the detected bounding boxes need to be to the ground-truth bounding boxes, and how close the detected temporal duration of an action need to be to the actual duration in the ground truth. Other imperfections such as multiple detections for a single ground truth action can similarly be handled. An intuitive way to express *Recall* and *Precision* in terms of matched detections is as follows:

$$\begin{aligned} \text{Recall}(\boldsymbol{G},\boldsymbol{D}) &= \frac{\sum_v \sum_a 1_{G^{v,a}\text{finds match in } D^v}}{\sum_v |G^v|} \\ \text{Precision}(\boldsymbol{G},\boldsymbol{D}) &= \frac{\sum_v \sum_a 1_{D^{v,a}\text{finds match in } G^v}}{\sum_v |D^v|} \end{aligned} \tag{3}$$

where 1_ω is the indicator function returning 1 if condition ω holds and 0 else; v is a video index and a an activity index.

Notice that both measures search for a match in a corresponding action list: *Recall* requires matching of each action in the ground truth to one of the actions in the detection list, whereas *Precision* requires matching of each action in the detected list to one of the actions in the ground-truth list. This is done in two steps by defining first the two functions $\beta(a, S)$ and $\Upsilon(g, d)$:

- For a given action a, $\beta(a, S)$ gives the best match in the set S of actions, which can be detected actions or ground-truth actions;
- For a pair of ground-truth action g and detected action d, $\Upsilon(g, d)$ determines whether the match between g and d satisfies our criteria on geometric and temporal overlaps. At this stage Υ can veto a match, if it is of poor quality.

The definitions in (3) can thus be refined as follows:

$$\begin{aligned}\text{Recall}(\boldsymbol{G}, \boldsymbol{D}) &= \frac{\sum_v \sum_a 1_{\Upsilon(G^{v,a}, \beta(G^{v,a}, D^v))}}{\sum_v |G^v|} \\ \text{Precision}(\boldsymbol{G}, \boldsymbol{D}) &= \frac{\sum_v \sum_a 1_{\Upsilon(\beta(D^{v,a}, G^v), D^{v,a}))}}{\sum_v |D^v|}\end{aligned} \tag{4}$$

Qualifying the best match $\beta(a, S)$ is done by maximizing the normalized overlap $\mathcal{O}$ between two actions a and b over all respective frames, where $\mathcal{O}$ is defined as the Sørensen-Dice coefficient:

$$\mathcal{O}(a, b) = \begin{cases} \frac{2 \cdot Area(a \cap b)}{Area(a) + Area(b)} & \text{if } Class(a) = Class(b) \\ 0 & \text{else} \end{cases} \tag{5}$$

Here $Area(a)$ is the sum of the areas of the bounding boxes of action a and $\cap$ is the intersection operator returning the overlap of two actions. The overlap is calculated framewise and summed over all frames.

More formally, for any given video v, the following two conditions hold:

$$\begin{aligned}\forall a, a' \in G^v \times G^v : &\beta(G^{v,a}, D^v) \neq \beta(G^{v,a'}, D^v) \\ \forall a, a' \in D^v \times D^v : &\beta(D^{v,a}, G^v) \neq \beta(D^{v,a'}, G^v)\end{aligned} \tag{6}$$

As a consequence, calculating $\beta(a, S)$ maximizes normalized overlap $\mathcal{O}$ as defined in (5) subject to constraints (6). These constraints preclude the matching of a single groundtruth action to multiple detected actions, and vice-versa. This maximization is made in a greedy way: $\mathcal{O}$ is calculated for all possible pairs $(G^{v,a}, D^{v,a'})$ and then the maximum value is searched, and the assignment chosen for the corresponding actions a and a'. These actions are then removed from respective lists, and the algorithm proceeds iteratively searching for the next best match, since the video may contain more than one action type.

$\Upsilon(g, d)$ decides whether a pair of ground truth action g and detected action d are *sufficiently* matched based on four criteria, two of which are spatial and two are temporal. Here we have used a simplified notation by denoting the ground-truth action as $g = G^{v,a}$ and the detected action as $d = D^{v,a'}$. We first describe these criteria intuitively and then formalize them in equation (7).

A detected action d can be matched to a ground truth action g if all of the following criteria are satisfied :

Sufficient temporal frame-wise recall — the number of frames which are part of both actions is above an adequate proportion of the number frames, i.e., a sufficiently long duration of the action has been correctly found;

Sufficient temporal frame-wise precision — the number of frames which are part of both actions is above an adequate proportion of the number frames in the detected set, i.e., the detected excess duration is small enough;

Sufficient spatial pixel-wise recall — the size of the common areas between the bounding boxes is large enough with respect to the size of the bounding boxes in the ground-truth set, i.e., a sufficiently large portion of the ground truth rectangles is correctly found. In order to ignore temporal differences, this calculation is done frame wise and includes only frames which are part of both actions, d and g;

Sufficient spatial pixel-wise precision — the size of common areas between the bounding boxes is large enough with respect to the size of the bounding boxes in the detected set, i.e., the space detected in excess is sufficiently small. In order to ignore temporal differences, this calculation is done frame wise and includes only frames which are part of both actions, d and g;

Correct classification — d and g have the same action class.

We denote by $d|_g$ the set of bounding boxes of the detection action d restricted to the frames which are also part of ground-truth action g. Then, the above criteria can be expressed as

$$\Upsilon(g,d) = \left\{ \begin{array}{ll} \frac{Area(g \cap d)}{Area(g|_d)} > t_{sr} & \text{and} \\ \frac{Area(g \cap d)}{Area(d|_g)} > t_{sp} & \text{and} \\ \frac{NoFrames(g \cap d)}{NoFrames(g)} > t_{tr} & \text{and} \\ \frac{NoFrames(g \cap d)}{NoFrames(d)} > t_{tp} & \text{and} \\ Class(g) = Class(d) & \end{array} \right\} \quad (7)$$

where $NoFrames(a)$ is the number of frames in set a. The decision whether the two actions g and d are correctly matched depends therefore on the threshold values $t_{sr}, t_{sp}, t_{tr}, t_{tp}$, which threshold, respectively, spatial pixel-wise recall, spatial pixel-wise precision, temporal frame-wise recall, and temporal frame-wise precision.

2.2. *Quantity/Quality plots*

We had put forward the necessity to consider the quality of detection with respect to the quantity of detection, as an inherent property of any method to assess algorithms. In our work, the quantity-quality interrelationship manifests itself through the dependence of *Recall* and *Precision* on the thresholds t_{sr}, t_{sp}, t_{tr} and t_{tp}. For this reason, an integral part of the proposed performance evaluation framework is a set of graphs which illustrate this dependence, similar to the graphs proposed in [52].

For each algorithm to be assessed, a number of diagrams are created, each one showing the performance as a function of one of the quality measures, that is, dependence on one of the thresholds. The performance graphs are produced by varying one threshold (assigned to the x-axis) in the interval $[0, 1]$, while the other three thresholds are kept at fixed lowest reasonable values, and plotting *Recall* and *Precision* and *F-Score* on the y-axis of the graphs. This results in 4 graphs containing each 3 curves.

Figures 7 and 8 in the experimental part (section 4) show examples of graphs obtained this way. These can be easily interpreted as recognition performance versus detection quality curves. Section 4.4 gives a more details on how to read these diagrams based on examples of actual detection methods.

Figure 2a shows a toy example involving an action of type *Discussion* covering a single frame. The ground truth bounding box, in blue, is labeled "1". Two different detection methods have been applied resulting in two bounding boxes, in red, and labeled "2" and "3", respectively. Since we are dealing with a single frame, temporal thresholds cannot be applied. The graphs for varying spatial thresholds are shown in Figures 2b and 2c, respectively. In this simple example, *Recall*, *Precision* and *F-Score* graphs collapse into one, since the measures are 1 if the bounding box is considered as detected, and 0 otherwise.

Since the bounding box "2" resulting from one of the methods is completely included in the ground-truth bounding box, varying the threshold t_{sp} (the required spatial pixel-wise precision) does not change the result: in fact, even for t_{sp}=1, which does not allow for any nonground-truth pixels, will consider the bounding box to be correctly detected. Varying t_{sr} (the required spatial pixel-wise recall), on the other hand, results in a drop of performance at roughly $t_{sr} = 0.25$. In order words, once we require more than 1/4 of the ground truth bounding box to be detected, the bounding box "2" is not considered as detected anymore. This corresponds to what we observe in Figure 2a.

Bounding box "3", produced by the other method, has only a small overlap with the ground truth-bounding box. Only a small part of the ground truth bounding box is detected, and in addition algorithm "3" produces a large spurious area. Consequently, we see an early drop in performance varying any one of the thresholds, $t_{sr} = 1$ or $t_{sp} = 1$.

2.3. *Ranking*

Ranking a set of detection algorithms according to a single performance measure should minimize the dependence on external parameters of the performance metric. This can be achieved by basing the final performance measure on an integration of the *F-Score* measure on the whole range of possible threshold values. In particular, four measures are created, each one measuring the performance while varying one of the thresholds while keeping the other ones fixed at a very low value (generally $\epsilon = 0.1$). In the following, we denote by $F(t_{sr}, t_{sp}, t_{tr}, t_{tp})$ the *F-Score* of equation (2) depending on the quality constraints. We get 4 integrals, each one showing the performance sensitivity averaged over the range of thresholds while three remaining variables (in both *Precision* or *Recall*) must satisfy a minimum of quality level, where the threshold is set to a reasonably small value (we choose ϵ=0.1). Thus,

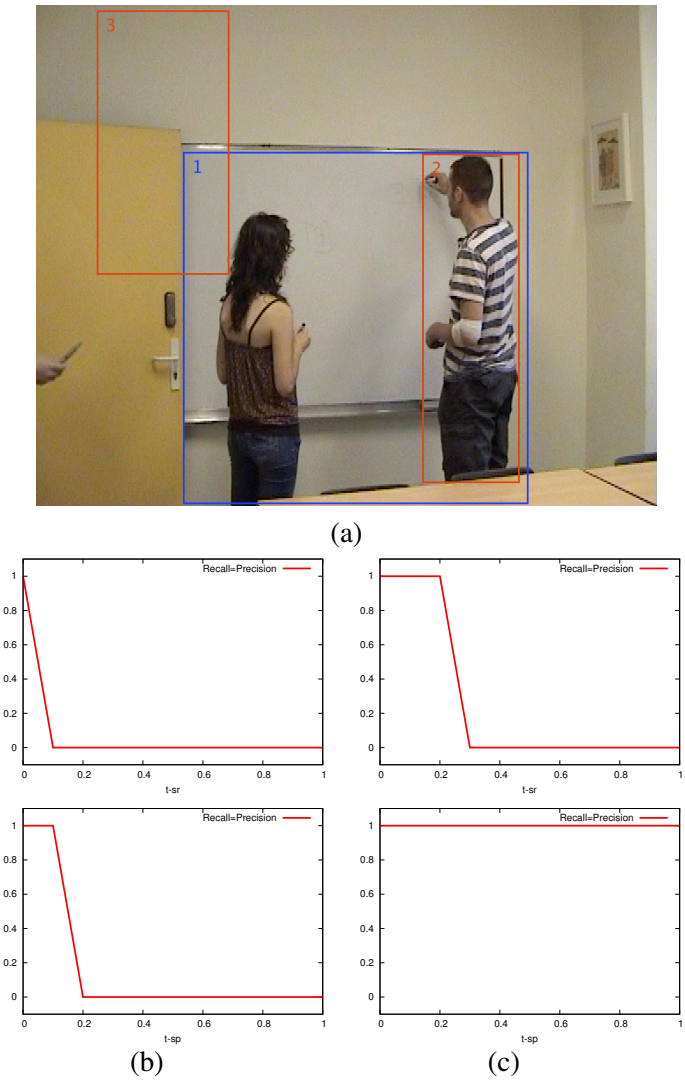


Figure 2: An example of evaluation plots for a single frame taken from the ICPR HARL competition: (a) a single frame with ground truth bounding box (labeled "1", in blue) and two bounding boxes corresponding, respectively, to two respective methods (labeled "2" and "3", in red); (b) evaluation curves for the method, labeled "3"); (c) evaluation curves for the method, labeled "2").

we get:

$$\begin{aligned} I_{sr} &= \int_0^1 F(u_{sr}, t_{sp}, t_{tr}, t_{tp})\, du_{sr} \\ I_{sp} &= \int_0^1 F(t_{sr}, u_{sp}, t_{tr}, t_{tp})\, du_{sp} \\ I_{tr} &= \int_0^1 F(t_{sr}, t_{sp}, u_{tr}, t_{tp})\, du_{tr} \\ I_{tp} &= \int_0^1 F(t_{sr}, t_{sp}, t_{tr}, u_{tp})\, du_{tp} \end{aligned} \tag{8}$$

In practice, we sample the *Precision* and *Recall* values in small steps and find the integrals numerically. The final value used for ranking is the mean over these four values:

$$\text{IntegratedPerformance} = \frac{1}{4}(I_{sr} + I_{sp} + I_{tr} + I_{tp}) \tag{9}$$

This integrated performance measure relates to the areas under the curves in the graphs described in section 2.2. In section 4 it will be experimentally shown, that this measure is quite invariant to changes in annotation styles.

Confusion matrices

The goal of the proposed performance metric is to go beyond classification, as the evaluated vision tasks also require detection and localization. However, it might be interesting to complete the traditional precision and recall measures with a confusion matrix which illustrates the pure classification performance of the evaluated methods. This can be done easily by associating a detected action to each ground truth rectangle using equations (5) and (7), while removing the class equality constraint from (7). The pairs ground truth — detected actions can be used to calculate a confusion matrix (see figure 10 for examples).

Note that the confusion matrix ignores actions which have not been detected, actions with no decision outcome. Therefore, unlike in classification tasks, the recognition rate (accuracy) cannot be determined from its diagonal. For this reason the confusion matrix must be accompanied by precision and recall values.

3. The LIRIS / ICPR 2012 HARL dataset

The LIRIS human activities dataset has been designed for recognizing complex and realistic actions in a set of videos, where each video may contain one or more actions concurrently. Table 1 shows the list of actions to be recognized. Some of them are interactions between two or more humans, like *discussion*,

DI	Discussion of two or several people	HH
GI	A person gives an item to a second person	HH,HO
BO	An item is picked up or put down (into/from a box, drawer, desk etc.)	HO
EN	A person enters or leaves an room	-
ET	A person tries to enter unsuccessfully	-
LO	A person unlocks a room and then enters	-
UB	A person leaves baggage unattended	HO
HS	Handshaking of two people	HH
KB	A person types on a keyboard	HO
TE	A person talks on a telephone	HO

Table 1: The behavior classes in the dataset. Some of the actions are human-human interactions (HH) or human-object interactions (HO).

giving an item etc. Other actions are characterized as interactions between humans and objects, for instance *talking on a telephone*, *leaving baggage unattended* etc. Note that simple "actions" as walking and running are not part of the events to be detected. The dataset therefore contains motion which is not necessarily relevant for the tasks at hand.

The dataset is publicly available online on a dedicated web site[3]. It is organized into two different and independent sets, shot with two different cameras:

D1/robot-kinect The videos of this set have been shot using a mobile robot of model Pekee II, shown in figure 3. During the tests the robot was controlled manually through a joystick. It was equipped with a consumer depth camera of type Primesense/MS Kinect, which delivers color images as well as 11bit depth images, both at a spatial resolution of 640×480 pixels, at 25 frames per second (see figures 4a—c). In the proposed dataset the RGB information has been converted to grayscale.

The Kinect module has been calibrated; the calibration information and its software are provided allowing users to calculate the coordinates in the grayscale image for each pixel of the corresponding depth image.

D2/fixed-camcorder The videos of this set have been shot with a consumer camcorder (a Sony DCR-HC51) mounted on a tripod. The camera was fixed (zero ego-motion), the videos have been shot in a spatial resolution of 720×576 pixels at 25 frames per second (see figure 4d).

[3] http://liris.cnrs.fr/voir/activities-dataset

Figure 3: The Pekee II mobile robot in our setup with the Kinect module during the shooting of the dataset.

The two sets D1 and D2 are *NOT* completely independent, as most of the D2 videos are shots from the same scenes captured in D1 but taken from a different viewpoint.

Care has been taken to ensure that the dataset is as realistic as possible:

- The actions have been performed by a group of 21 different people.
- The actions have been shot from various viewpoints and different settings to avoid the possibility of learning actions from background features.
- Correlation between camera motion and activities has been avoided.

In order to make the dataset more challenging than previous datasets, the actions are less focused on low-level characteristics and defined more by semantics and context:

- The *discussion* action can take place anywhere, either by people standing in some room or in an aisle without any support, or in front of a whiteboard or blackboard, or by people sitting on chairs.
- The action *enter or leave a room* can involve opening a door and passing through or passing through an already open door.

- Three actions involve very similar motion, the difference being the context : *entering a room*, *unlocking a door and then entering a room* and *trying to enter a room without being able to open the door*.
- The action of *an item being picked up or put down (into/from a box, drawer, desk etc.)* is very similar to the action of *a person leaving a baggage unattended (drop and leave)*, as both involve very similar human-object interactions. The difference is mainly defined through the context.
- We took care to use different telephones in the action *telephone conversation*: classical office telephones, cell phones, wall mounted phones.
- Actions like *handshaking* and *giving an item* can occur before, after or in the middle of other actions like *discussion*, *typing on a keyboard* etc.

The acquisition conditions have *not* been artificially improved, which means that the following additional difficulties are present in the dataset :

- Non-uniform lighting and lighting changes when doors open and close
- The Kinect camera's gain control is rather slow compared to other cameras. This is not the case for the Sony camcorder.
- The depth data delivered by the Kinect camera is disturbed by transparencies like windows etc. This is due to the data acquisition method (shape from structured light).
- The data taken with the mobile robot is subject to vibrations when the robot accelerates or slows down. This reflects realistic conditions in a mobile robotics environment.

The full data set contains 828 actions (subsets D1 and D2) by 21 different people. Each video may contain one or several people performing one or several actions. Example images for the different activity classes are shown in figure 5.

All actions are localized in time and space, and ground truth bounding boxes are provided. Figure 6 shows a frame with annotated bounding boxes in a screen shot of the annotation/viewing tool provided with the dataset. Each video has been annotated by one of 10 annotators, and then verified by a different annotator to keep the annotations as coherent as possible.

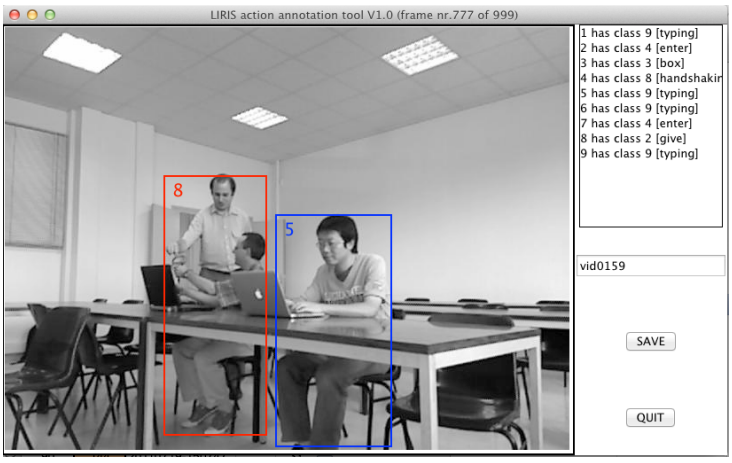


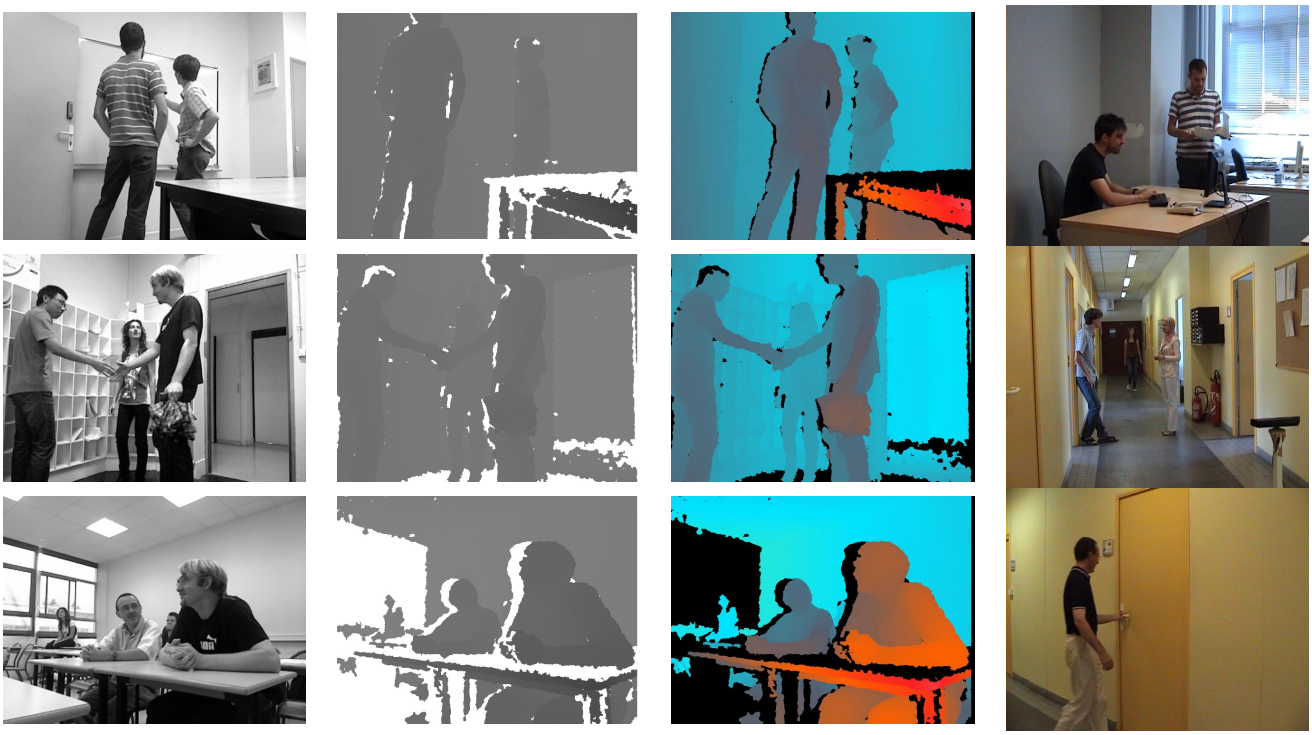
Figure 6: The videos are annotated: each action is localized through a set of bounding boxes over a contiguous sequence of frames.

4. Results of the ICPR 2012 HARL competition

The proposed performance metric was tested on six different detection and recognition algorithms. Four methods correspond to submissions of the ICPR 2012 HARL competition, which was held in conjunction with the *International Conference on Pattern Recognition 2012*. Two additional methods have been applied to the same dataset.

The HARL competition took place during roughly 12 months from October 2011 to October 2012. The video frames of the competition dataset (described in section 3) were published in October 2011 and the ground-truth annotations of the training set were released in December 2011. The participants had 7 months to develop and train their system. In mid July 2012 annotations of the test set were published and in September 2012 the results had to be submitted to the competition committee. A special session dedicated to the HARL competition was held during the ICPR conference in November 2012. The competition has attracted great interest: 70 teams from all over the world registered to the competition and downloaded the dataset, which appeared to be more difficult than existing datasets at that time, as anticipated. The task of not only classifying, but also locating activities in space and in time is still a hard one. Four teams finally managed to solve the problem and to submit their results.

We distinguished the six methods in the following way: the four participations were identified by their participation numbers (13, 49, 51, 59), and the two additional methods were identified by letters A and B.



(a) (b) (c) (d)

Figure 4: The dataset has been shot with two different cameras, Kinect camera and a color camera. (a) Kinect - grayscale image; (b) Kinect - depth image; (c) Kinect - color coded depth image; (d) color images from the Sony camcorder.

4.1. Evaluated methods

HARL participant No. 13 (subset D2)
Participating team No. 13 came from Spain and submitted a run for dataset D2 (Sony color frames):

Juan C. SanMiguel and *Sergio Suja*
Video Processing and Understanding Lab
Universidad Autonoma of Madrid, Spain

The description of the submitted method is yet undisclosed for reasons related to protection of intellectual property rights.

HARL participant No. 49 (subset D1)
Participating team No. 49 was a collaboration between institutions in Singapore and a US university, and they submitted a run for dataset D1 (Kinect frames):

Bingbing Ni and *Yong Pei*
Advanced Digital Sciences Center, Singapore
Jun Tan, Jian Dong and *Shuicheng Yan*
National University of Singapore, Singapore
Pierre Moulin
University of Illinois at Urbana-Champaign

The submitted method (published in [54]) uses low-level features and mid-level features calculated on detected and tracked people using [55] and detected objects specific to the dataset (doors, mailboxes etc.). The following features were calculated: (i) pose and appearance information on people and on objects; (ii) geometric contextual attributes on pairs of neighboring items, where each item may be a person or an object; (iii) scene type attributes obtained by clustering depth histograms. These heterogeneous features were integrated through a Bayesian network learned from the training set.

HARL Participant No. 51 (subset D2)
Participating team No. 51 came from China and submitted a run for dataset D2 (Sony color frames):

Yonghao He, Hao Liu, Wei Sui, Shiming Xiang
and *Chunhong Pan*
Institute of Automation,
Chinese Academy of Sciences, Beijing

The submitted method is an adaptation of existing work in [56]. Space-time interest points were extracted and HoG-HoF (histogram of gradients and histogram of motion flow) features were extracted from the local patches [57]. Then 10

Figure 5: Example frames for various activity classes (D1/Kinect grayscale shown only)

one-against-all SVM classifiers were trained for the 10 activity categories. Activities were detected and localized by shifting sub-volumes over the video and maximizing mutual information, as in [56]. Adaptations concerned reduction of search boundaries according to maximize mutual information and calculations of proper step widths, which significantly increased performance compared with the original brute force search.

HARL participant No. 59 (subset D1)

Participating team No. 59 came from India and was an academic / industrial collaboration. They submitted a run for dataset D2 (Sony color frames):

Tanushyam Chattopadhyay, Sangheeta Roy and *Aniruddha Sinha*
Innovation Lab, Tata Consultancy Services, Kolkata
Dipti Prasad Mukherjee and *Apurbaa Mallik*
Indian Statistical Institute, Kolkata

The submitted method detects and classifies actions in the dataset, but does not localize them. It segments and extracts "interesting" moving objects in the scene based on motion and entropy. HoF features were calculated according to [58] in a hierarchical way using a pyramid.

Method A - (subset D1)

To create a baseline, we calculated 3D (depth) features proposed in [59] and extracted them with dense-sampling on sliding cuboids. Bounding boxes were estimated on the test set using the same pre-processing step as method B [60]: people were tracked using the Dalal/Triggs detector [55], and candidate bounding boxes were created based on pairwise combinations of tracklets. From the acquired features, codebooks trained through k-means clustering and videos were represented as bags of words (BoW) on space-time sliding windows. For the recognition part an SVM classifier was trained. We detected activities including a *no-action* class into the classifier which was trained trough boot-strapping.

Method B - (subset D1)

Method B (published in [60]) shares the features and the preprocessing steps as the winning entry of the HARL competition, entry nr. 49. In particular, tracklets are created by the Dalal/Triggs detector and combined into larger bounding boxes. Instead of learning a belief network, each activity is modelled as a deformable parts model in the spirit of [61] and learned using structured SVM.

No.	Set	Recall	Precision	F-Score
49	D1	0.74	0.41	**0.53**
59	D1	0.08	0.17	0.11
A	D1	0.34	0.24	0.28
B	D1	0.74	0.41	**0.53**
13	D2	0.35	0.66	**0.46**
51	D2	0.30	0.46	0.36

Table 2: Results **without localization**. The bounding boxes of the annotation are not used.

No.	Set	Recall	Precision	F-Score
49	D1	0.63	0.33	0.44
59	D1	N/A	N/A	N/A
A	D1	0.27	0.18	0.22
B	D1	0.67	0.36	0.47
13	D2	0.04	0.08	**0.05**
51	D2	0.03	0.04	0.03

Table 3: Results with **fixed quality constraints**: all thresholds are set to 0.1. No localization information has been submitted for method No. 59.

4.2. Results without localization

Table 2 shows a preliminary evaluation of *Recall* and *Precision* values calculated according to equations (1) but without any localization information, i.e., ignoring the bounding box related information in the ground truth. Half of the participants submitted results for dataset D1 (kinect) and the other half submitted runs for dataset D2 (color frames). The two additional methods were applied to dataset D1. Results of the two datasets cannot be directly compared, of course. As expected, the obtained results are better on the Kinect data than on the color frames, since the depth data is richer in information as compared to color data for these scenes. We consider a recall rate of 74% as an excellent result for this difficult dataset with high intra-class variations. A precision value of 41% indicates that, roughly, for each correctly detected activity, a second incorrect activity has been detected. Note that no confusion matrices can be given if localization is not used. Methods 49 and B obtained the same performance in this setting, as the difference lies in the way how activities are localized.

4.3. Results with localization

Table 3 gives performance measures which do use localization information from ground truth and detection. For a first experiment, all quality thresholds have been fixed to a low value of

t_{sr}=t_{sp}=t_{tr}=t_{tp}=0.1. In other words, a ground-truth action g is matched to a detected action a if and only if

- at least 10% of the ground-truth frames are detected (t_{tr});
- at least 10% of the detected frames are also in the ground-truth (t_{tp});
- at least 10% of the pixels of the ground-truth bounding boxes have been detected, only counting frames which appear in both ground truth and detection (t_{sr});
- at least 10% of the pixels of the detected bounding boxes are also in the ground-truth bounding boxes, only counting frames which appear in both ground truth and detection (t_{sp}).

These conditions correspond to equation (7) in section 2.

When localization information is taken into account, differences in performance between the algorithms become much more evident. While the differences were modest when no penalty was applied on localization, now a clear winner emerges in competition participant No. 49, only slightly topped by not participating method B. Including these new constraints due to localization, *Recall* drops from 74% to 63%, and *Precision* drops from 41% to 33% for this participant. The performance for the winning entry, submitted for dataset D1, cannot be directly compared to the performance of methods No. 13 and No. 51, which were submitted for dataset D2. However, given the difference in F-scores, with 44% on one hand, and 5% and 3% on the other hand, it is safe to announce for a clear winner of the contest.

With an F-Score of 0.22, the baseline method based on a bags of words representation (method A) fares reasonable well compared to the winning methods based on more sophisticated methods integrating spatio-temporal relationships through belief nets (method 49, F-Score of 0.44) and deformable parts models (method B, F-Score of 47). Integrating spatial relationships is clearly important.

4.4. Dependence on quality

The measures described above have been calculated using thresholds set to 0.1, which seems to be a good compromise given the high spatial and temporal variations of human activities. However, interesting information on the behavior of a detection algorithm can be obtained by calculating *Recall* and *Precision* measures over varying thresholds and creating plots, as explained in section 2.2. Figures 7 and 8 show these graphs. Each column corresponds to a method, and each of the four rows corresponds to a situation where only one of the four thresholds t_{sr}, t_{sp}, t_{tr} and t_{tp} is varied from 0 to 1, while the other three thresholds are kept to a fixed value of 0.1.

Equation →		(8)	(8)	(8)	(8)	(9)
No.	Set	I_{tr}	I_{tp}	I_{sr}	I_{sp}	Total
49	D1	0.27	0.37	0.29	0.37	0.33
59	D1	N/A	N/A	N/A	N/A	N/A
A	D1	0.14	0.15	0.15	0.19	0.16
B	D1	0.30	0.38	0.32	0.40	0.35
13	D2	0.03	0.03	0.02	0.03	**0.03**
51	D2	0.03	0.00	0.01	0.01	0.02

Table 4: Results **integrated over all quality constraints**: for each column of type I_* a single threshold is varied and the others are fixed. The total is the mean value over these indicators. No localization information has been submitted for method No. 59.

Focusing our attention on the winning entry, method No. 49 shown in the rightmost column, we can deduce valuable information from the first diagram in the top row, where the threshold t_{tr} of the temporal frame-wise recall is varied. The highest performance is obtained for $t_{tr} = 0$: Recall=65% and Precision=36%. Note that the thresholding condition requires the thresholded quantity to be strictly larger than 0, as indicated in equation (7). When we increase the threshold in small steps up to t_{tr}=1, we can see an almost linear drop of the *Precision* and *Recall* measures. At the right end of the diagram we see that we obtain a performance of 8% and 5%, respectively, for t_{tr}=1, which gives the performance for the case where all frames of an activity need to be detected in order for the activity itself to be counted as detected.

A similar behavior can be seen when threshold t_{tp} (temporal frame-wise precision) is varied, illustrated in the second row of figure 7. At t_{tp}=1, when we require that not a single spurious frame outside ground-truth activity is detected, we still get performance measures of 30% and 16%, respectively.

The last two rows of figure 7 illustrate the behavior when spatial overlap is considered. Both diagrams show performance figures approaching zero when the respective threshold approaches 1. This shows that it is extremely rare for a ground truth activity to be consistently (spatially) included in the corresponding detected activity over all frames, as indicated in the third row, and it is extremely rare for a detected activity to be (spatially) included in the corresponding ground-truth activity over all frames, as indicated in the last row.

These indications of behavior over varying quality con-

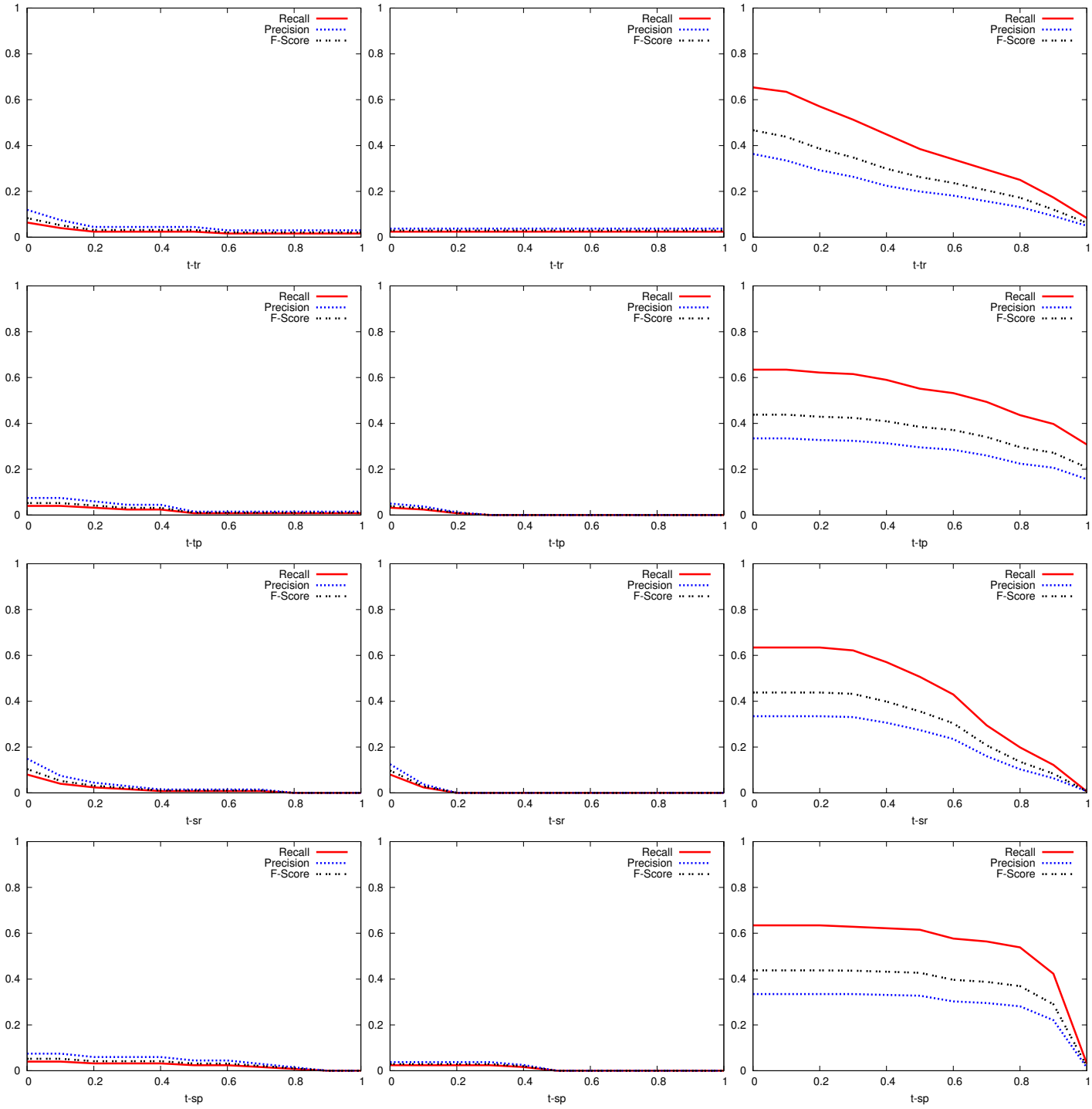


Figure 7: Performance curves for the three participating methods in view of localization data (OR, under the constraint of localization data). They are obtained by varying a single quality constraint and keeping the other ones at 0.1 level. From top to bottom, the following constraints are varied: temporal recall, temporal precision, spatial recall, spatial precision.

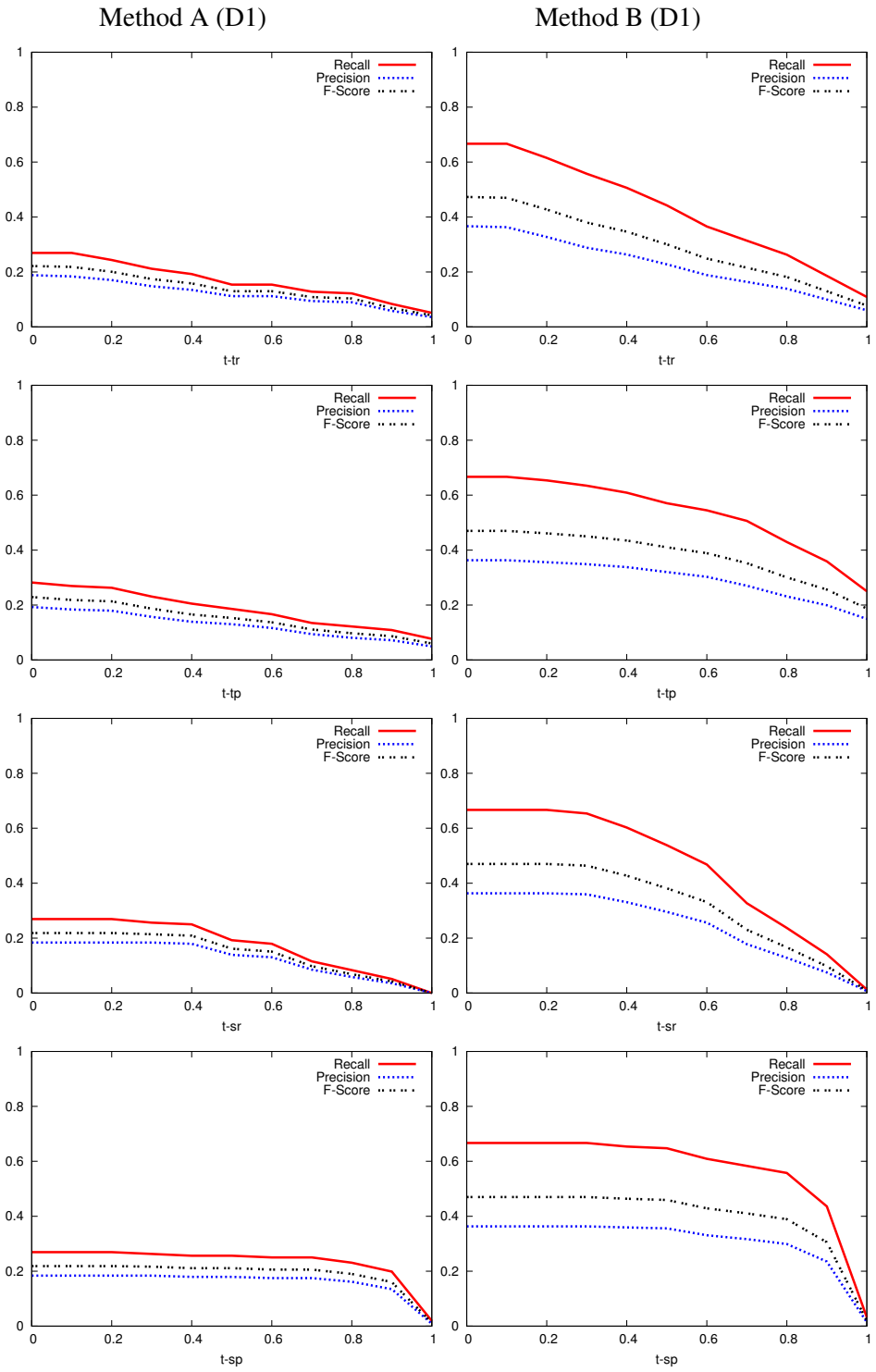


Figure 8: Performance curves for the two additional methods in view of localization data (OR, under the constraint of localization data). They are obtained by varying a single quality constraint and keeping the other ones at 0.1 level. From top to bottom, the following constraints are varied: temporal recall, temporal precision, spatial recall, spatial precision.

Annotator Id	N.o. runs	F-Score (fixed thresholds) t_*=0.1	t_*=0.5	t_*=0.8	Integrated performance
1	6	1.0	0.93	0.30	0.86
2	6	1.0	0.91	0.39	0.87
3	6	1.0	0.91	0.35	0.86
4	6	1.0	0.93	0.35	0.86
5	6	1.0	0.87	0.35	0.86
6	6	1.0	0.83	0.24	0.84
7	6	1.0	0.70	0.17	0.83
1-7	21	1.0	0.86	0.31	0.85

Table 5: Estimation of soft upper bounds on the proposed performance measures: mean performance values calculated on different ground-truth annotations (t* signifies all four thresholds collectively set at that the given value). In contrast to classical measures as in Eq. 7, integrated performance stays quite stable over different annotators.

straints can be captured in a performance measure, as given in equation (8) in section 2.3. Intuitively, the measures are means of F-Score over the threshold variations. They are given in figure 4 for the different participants of the competition.

4.5. Soft upper bounds on performance

The upper bound for any of the performance measures, (*Precision*, *Recall* and *F-Score*) is in principle 1. However, ground-truth annotations are subjective and inherently imprecise, so a totally precise localization resulting in *Recall=Precision=F-Score*=1 may not be expected for any method. It is therefore interesting to estimate "soft" upper bounds on the performance measures corresponding to the average agreement score of human annotators (inter-subject agreement), which is defined as expected value of the performance measures when different test subjects do the localization task. To estimate these bounds, we selected a subset of 9 videos containing 9 actions and had these actions annotated by 7 different people. From this pool of annotations, pairs of annotations were selected where the first one was used as ground truth and the second one as (virtual) detection.

Table 5 shows means of *F-Score* obtained for sets of such pairs. The last row shows the mean over all possible 21 combinations of pairs of this set of 7 annotators. The first seven rows give the different means for different annotators, where each row corresponds to a mean over 6 runs (one annotator against the 6 other annotators). The different columns correspond to *F-Score* for different quality constraints, as given in equation (2), as well as integrated performance, as given in equation (9).

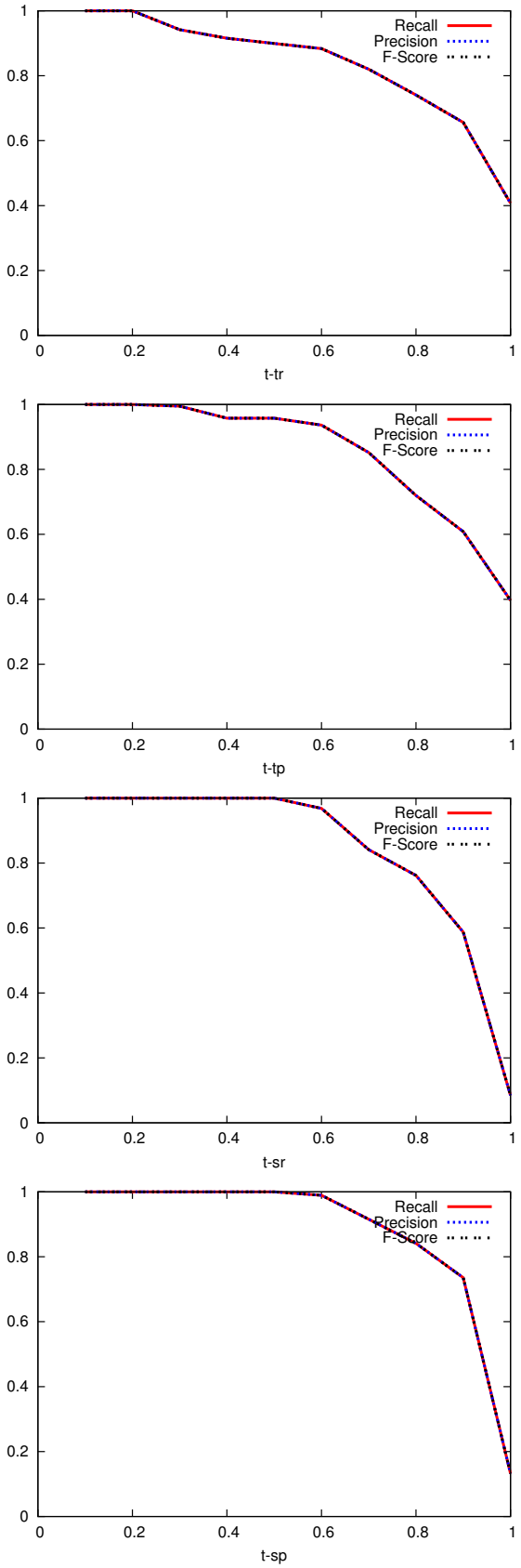


Figure 9: Estimation of soft upper bounds on the proposed performance measures: mean performance curves calculated on pairs of different ground truth annotations.

The given figures may seem quite low, especially for thresholds equal to 0.8. These low detection rates among different manual annotations can be explained by the fact that the threshold is enforced jointly in the temporal and the spatial domain. It is very difficult to create similar annotations in all aspects, i.e., not to cut away relevant parts or not to add irrelevant parts, both temporally and spatially. We can also see that while classical performance measures (measures that do not take into account the quality factors as in Eqs. 1-2) seem to vary across annotators, the proposed integrated performance measure stays quite stable over different annotators. We claim that this invariance to subtle changes in annotations is a major advantage of this new metric. This does not mean that the measure loses its discriminative power in comparing different methods, as can be seen in Table 4: for the winning entry, performance is measured as 33%, whereas the other two entries are measured as 3% and 2%, respectively. The large difference in the methods' performance is illustrated by the curves given in figures 7 and 8.

"Soft" upper bounds have also been calculated for the performance curves proposed in section 2.2. These curves are shown in figure 9, where each row corresponds to a variation in one of the four thresholds, in the same way as shown in figure 7. In figure 9, each point of a plot corresponds to a mean calculated over the 21 different values obtained by taking all possible pairwise combinations of annotations. Note that the plots of *Recall*, *Precision* and *F-Score* are identical for each diagram. This is due to the fact that all annotators have annotated the same activities and that no annotator added a new false activity. In other words, the differences in annotations are only in the coordinates and number of the bounding boxes, not in the number of found activities.

4.6. Confusion matrices

Figure 10 shows confusion matrices for the methods for which localization information (bounding boxes) has been submitted. These matrices contain complementary information to *Recall* and *Precision* values, but otherwise they cannot be used as indicators for detection performance of a method. In particular, calculating the difference between the ground-truth class and the detected class requires the assignment of a detected activity for each ground-truth activity, which can only be done through localization information, i.e., bounding boxes. Unmatched activities are not included in these matrices, which therefore lack any information on the amount of unfound activities or false alarms.

Considering the matrix for the winning entry, the bottom matrix in the figure, we see that the actions *Handshaking* and *Give Item* are frequently confused, which is not surprising given the similar motion involved in both actions. Activities *Telephone call* and *Pick up / Put down Object* are also sometimes confused, which may eventually be explained by the context model used in the method. Both methods take place in similar contexts, and picking up and putting down a telephone has been annotated as *Pick up / Put down Object*. Further confusions are *Discussion* and *Give Item*, which both involve a group of people standing close together and interacting.

4.7. Implementation and tools

An open-source implementation for Windows, Linux and Mac OS of the proposed performance metric is available online[4]. The software allows to calculate *Recall*, *Precision* and *F-Score* for fixed (selectable) thresholds as well as integrated performance, to plot and export performance curves and confusion matrices. Two versions are available: one with a graphical user interface and one version with a scriptable command line interface. It comes with software allowing to view ground-truth annotations superimposed on videos, as well as software which allows to create new annotations.

5. Conclusion

This paper has introduced a new performance metric which allows to evaluate human activity detection, recognition and localization algorithms. Taking into account localization information is a non-trivial task, as evaluation needs to decide for each activity whether it has been successfully detected based on detection quality constraints. The inherent dependency between performance and quality has been identified and a set of quantity / quality curves has been introduced to describe the detection and localization behavior of a computer vision algorithm.

The proposed integrated performance measure is a new way to compare and rank detection and localization methods. Its advantages are two-fold:

- the measure is independent of quality constraints on detection, i.e. it is independent of arbitrary thresholds on spatial and temporal overlap;

[4] `http://liris.cnrs.fr/voir/activities-dataset`

Participant Nr. 13 (subset D2)

	DI	GI	BO	EN	ET	LO	UB	HS	KB	TE
DI	50	0	0	0	0	0	0	50	0	0
GI	50	0	50	0	0	0	0	0	0	0
BO	0	0	80	0	0	0	0	0	0	0
EN	0	0	0	100	0	0	0	0	0	0
ET	33	0	0	0	67	0	0	0	0	0
LO	0	0	0	0	100	0	0	0	0	0
UB	0	0	100	0	0	0	0	0	0	0
HS	0	0	100	0	0	0	0	0	0	0
KB	0	0	67	33	0	0	0	0	0	0
TE	0	0	0	0	0	0	0	0	0	100

Method A (subset D1)

	DI	GI	BO	EN	ET	LO	UB	HS	KB	TE
DI	60	0	7	33	0	0	0	0	0	0
GI	20	0	20	40	0	20	0	0	0	0
BO	0	0	29	43	7	7	0	0	0	14
EN	0	0	0	83	6	11	0	0	0	0
ET	0	0	0	100	0	0	0	0	0	0
LO	0	0	25	0	50	25	0	0	0	0
UB	0	0	17	83	0	0	0	0	0	0
HS	6	0	0	71	12	0	0	12	0	0
KB	0	0	0	60	0	0	0	0	40	0
TE	0	0	0	71	0	0	0	0	14	14

Participant Nr. 51 (subset D2)

	DI	GI	BO	EN	ET	LO	UB	HS	KB	TE
DI	100	0	0	0	0	0	0	0	0	0
GI	100	0	0	0	0	0	0	0	0	0
BO	50	0	0	0	0	25	0	25	0	0
EN	0	0	0	89	0	11	0	0	0	0
ET	0	0	0	25	25	25	0	25	0	0
LO	0	0	0	100	0	0	0	0	0	0
UB	100	0	0	0	0	0	0	0	0	0
HS	0	0	14	57	0	14	0	14	0	0
KB	100	0	0	0	0	0	0	0	0	0
TE	33	0	0	33	0	33	0	0	0	0

Method B (subset D1)

	DI	GI	BO	EN	ET	LO	UB	HS	KB	TE
DI	93	7	0	0	0	0	0	0	0	0
GI	33	50	0	0	0	17	0	0	0	0
BO	0	8	69	0	0	0	0	0	15	8
EN	0	0	0	96	4	0	0	0	0	0
ET	0	0	0	0	89	11	0	0	0	0
LO	0	0	0	11	0	89	0	0	0	0
UB	20	0	40	0	0	0	40	0	0	0
HS	6	59	6	6	0	0	0	24	0	0
KB	0	20	10	0	0	0	0	0	70	0
TE	0	0	57	0	0	0	0	0	14	29

Participant Nr. 49 (subset D1)

	DI	GI	BO	EN	ET	LO	UB	HS	KB	TE
DI	93	7	0	0	0	0	0	0	0	0
GI	33	50	0	0	0	17	0	0	0	0
BO	0	8	69	0	0	0	0	0	15	8
EN	0	0	0	96	4	0	0	0	0	0
ET	0	0	0	0	89	11	0	0	0	0
LO	0	0	0	11	0	89	0	0	0	0
UB	20	0	40	0	0	0	40	0	0	0
HS	6	56	6	6	0	0	0	25	0	0
KB	0	20	10	0	0	0	0	0	70	0
TE	0	0	57	0	0	0	0	0	14	29

Figure 10: Confusion matrices for the five methods providing localization data. They are calculated on actions satisfying quality constraints only. Rows are ground-truth classes, columns are detected classes.

- experiments described in this paper have shown, that the measure is less sensitive to annotator variance than the classical measures while at the same time allowing to discriminate between changes in performance of the algorithms.

The paper also describes the LIRIS human activities dataset as a new standard dataset, which allows to benchmark activity recognition algorithms based on realistic and difficult data. The proposed performance metric has been tested on the LIRIS dataset and on the detection methods submitted to the ICPR HARL 2012 competition.

References

[1] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: International Conference on Pattern Recognition, volume 3, 2004, pp. 32–36 Vol.3.

[2] L. Zelnik-Manor, M. Irani, Weizmann event-based analysis of video, 2013. URL: `http://www.wisdom.weizmann.ac.il/~vision/VideoAnalysis/Demos/EventDetection/EventDetection.html`.

[3] I. Laptev, Irisa download data/software, 2013. URL: `http://www.irisa.fr/vista/Equipe/People/Laptev/download.html`.

[4] I. Laptev, Hollywood2: Human actions and scenes dataset, 2013. URL: `http://www.irisa.fr/vista/actions/hollywood2/`.

[5] S. University, Olympic sports dataset, 2013. URL: `http://vision.stanford.edu/Datasets/OlympicSports/`.

[6] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, M. Boonstra, V. Korzhova, Performance Evaluation Protocol for Text, Face, Hands, Person and Vehicle Detection & Tracking in Video Analysis and Content Extraction (VACE-II), Technical Report, University of South Florida, Tampa, FL, USA, 2005.

[7] R. Collins, X. Zhou, S. Teh, An open source tracking testbed and evaluation web site, in: International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), Beijing, China, 2005, pp. 17–24.

[8] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, J. Zhang, Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 319–336.

[9] A. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and trecvid, in: International Workshop on Multimedia Information Retrieval, Santa Barbara, CA, USA, 2006, pp. 321–330.

[10] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, C. Rochet, The chil audiovisual corpus for lecture and meeting analysis inside smart rooms, Language Resources and Evaluation 41 (2007) 389–407.

[11] X. Xu, J. Tang, X. Zhang, X. Liu, H. Zhang, Y. Qiu, Exploring techniques for vision based human activity recognition: Methods, systems, and evaluation, Sensors 13(2) (2013) 1635–1650.

[12] O. Kliper-Gross, T. Hassner, L. Wolf, The action similarity labeling challenge, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2012) 615–621.

[13] J.A.Ward, P. Lukowicz, G. Tröster, Evaluating performance in continuous context recognition using event-driven error characterisation, in: Location- and Context-Awareness, LNCS, volume 3987, 2006, pp. 239–255.

[14] J. Ward, P. Lukowicz, G. Troster, T. Starner, Activity recognition of assembly tasks using body-worn microphones and accelerometers, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 1553–1567.

[15] J. A. Ward, P. Lukowicz, H. W. Gellersen, Performance metrics for activity recognition, ACM Transactions on Intelligent Systems Technology 2 (2011) 6:1–6:23.

[16] D. Minnen, T. Westeyn, T.Starne, J. Ward, P. Lukowicz, Performance metrics and evaluation issues for continuous activity recognition, Performance Metrics for Intelligent Systems (2006).

[17] H. A. T.L.M. van Kasteren, C. Ersoy, Effective performance metrics for evaluating activity recognition methods, in: ARCS Workshop on Context-Systems Design, Evaluation and Optimisation, 2011.

[18] J. Chaquet, E. Carmona, A. Fernandez-Caballero, A survey of video datasets for human action and activity recognition, Computer Vision and Image Understanding (2013) (to appear).

[19] Caviar: Context aware vision using image-based active recognition, 2013. URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm.

[20] INRIA, Etiseo video understanding evaluation, 2013. URL: http://www-sop.inria.fr/orion/ETISEO/index.htm.

[21] D. Tran, A. Sorokin, D. Forsyth, Human activity recognition with metric learning, 2013. URL: http://vision.cs.uiuc.edu/projects/activity/.

[22] J. Yuan, Z. Liu, Y. Wu, Discriminative video pattern search for efficient action detection, 2013. URL: http://users.eecs.northwestern.edu/~jyu410/index_files/actiondetection.html.

[23] R. Fisher, Behave: Computer-assisted prescreening of video streams for unusual activities, 2013. URL: http://homepages.inf.ed.ac.uk/rbf/BEHAVE/.

[24] C. for Biometrics, S. Research, Casia action database for recognition, 2013. URL: http://www.cbsr.ia.ac.cn/english/Action$\%$20Databases$\%$20EN.asp.

[25] U. of Surrey, CERTH-ITI, i3dpost multi-view human action datasets, 2013. URL: http://kahlan.eps.surrey.ac.uk/i3dpost_action/.

[26] V. G. Group, Tv humacn interactions dataset, 2013. URL: http://www.robots.ox.ac.uk/~vgg/data/tv_human_interactions/index.html.

[27] M. S. Ryoo, J. K. Aggarwal, Ut-interaction dataset, icpr contest on semantic description of human activities (sdha), 2013. URL: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.

[28] V. C. Group, Videoweb dataset, 2013. URL: http://www.ee.ucr.edu/~amitrc/vwdata.php.

[29] S. o. S. E. Reading University Computational Vision Group, Pets 2009 benchmark data, 2013. URL: http://www.cvg.rdg.ac.uk/PETS2009/a.html.

[30] S. S. W. Choi, K. Shahid, What are they doing? : Collective activity classification using spatio-temporal relationship among people., in: International ICCV Workshop on Visual Surveillance, 2009.

[31] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: International Conference on Computer Vision, 2009.

[32] M. Tenorth, J. Bandouch, M. Beetz, The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition, in: International ICCV Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences, 2009.

[33] INRIA, Inria xmas motion acquisition sequences (ixmas), 2013. URL: http://4drepository.inrialpes.fr/public/viewgroup/6.

[34] K. University, Muhavi: Multicamera human action video data, 2013. URL: http://dipersec.king.ac.uk/MuHAVi-MAS/.

[35] U. of Central Florida, Ucf aerial camera, rooftop camera and ground camera dataset, 2013. URL: http://vision.eecs.ucf.edu/data/UCF-ARG.html.

[36] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, J. Rousseau, Multiple cameras fall dataset, Technical Report 1350, DIRO-Université de Montréal, 2010.

[37] U. of Central Florida, Ucf aerial action dataset, 2013. URL: http://server.cs.ucf.edu/~vision/aerial/index.html.

[38] S. Lab., Hmdb: A large video database for human motion recognition, 2013. URL: http://serre-lab.clps.brown.edu/resources/HMDB/.

[39] U. of Central Florida, Ucf youtube action dataset, 2013. URL: http://www.cs.ucf.edu/~liujg/YouTube_Action_dataset.html.

[40] U. of Central Florida, Ucf sports action dataset, 2013. URL: http://crcv.ucf.edu/data/UCF_Sports_Action.php.

[41] A. R. Z. K. Soomro, M. Shah, A dataset of 101 human action classes from videos in the wild, in: THUMOS: The First International Workshop on Action Recognition with a Large Number of Classes, in conjunction with ICCV 2013, 2013. URL: http://crcv.ucf.edu/ICCV13-Action-Workshop/.

[42] Microsoft, Msr action recognition datasets and codes, 2013. URL: http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm.

[43] S. Fothergill, H. M. Mentis, P. Kohli, S. Nowozin, Instructing people for training gestural interactive systems, in: J. A. Konstan, E. H. Chi, K. Höök (Eds.), ACM Conference on Computer-Human Interaction, 2012, pp. 1737–1746.

[44] C. ChaLearn, Chalearn gesture dataset (cgd2011), 2013. URL: http://gesture.chalearn.org/data.

[45] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Berkeley mhad: A comprehensive multimodal human action database, in: International Workshop on Applications on Computer Vision, 2013.

[46] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in: 3rd International Workshop on Human Activity Understanding from 3D data (HAU3D-CVPRW), in conjunction with CVPR, 2013. URL: http://www.micc.unifi.it/vim/datasets/3dactions/.

[47] J. Sung, H. Koppula, B. Selman, A. Saxena, Cornell activity datasets: Cad-60 & cad-120, 2013. URL: http://pr.cs.cornell.edu/humanactivities/data.php.

[48] B. Ni, G. Wang, P. Moulin, Rgbd-hudaact: A color-depth video database for human daily activity recognition, in: International ICCV Workshop on Consumer Depth Cameras for Computer Vision, 2011. URL: http://www.adsc.illinois.edu/demos.html.

[49] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, D. Samaras, in: 3rd International Workshop on Human Activity Understanding from 3D Data (HAU3D-CVPRW), in conjunction with CVPR, 2013. URL: http://www.cs.stonybrook.edu/~kyun/research/kinect_interaction/index.html.

[50] A. Kläser, M. Marszałek, C. Schmid, A. Zisserman, Human focused action localization in video, in: International Workshop on Sign, Gesture, and Activity (SGA) in Conjunction with ECCV, 2010. URL: http://lear.inrialpes.fr/pubs/2010/KMSZ10.

[51] I. Laptev, P. Perez, Retrieving actions in movies, in: ICCV, 2007.

[52] C. Wolf, J.-M. Jolion, Object count/Area Graphs for the Evaluation of Object Detection and Segmentation Algorithms, International Journal on Document Analysis and Recognition 8 (2006) 280–296.

[53] C. van Rijsbergen, Information Retrieval, 2nd ed., Butterworths, London, 1979.

[54] B. Ni, Y. Pei, P. Moulin, Integrating multi stage depth induced contextual information for human action recognition and localization, in: International Conference on Automatic Face and Gesture Recognition, 2013.

[55] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, volume 1, 2005, pp. 886–893.

[56] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: International Conference on Computer Vision and Pattern Recognition, 2009.

[57] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[58] S. Mukherjee, S. Biswas, D. Mukherjee, Recognizing human action at a distance in video by key poses, IEEE Transactions on Circuits and Systems for Video Technology 21 (2011) 1228–1241.

[59] A. Kläser, M. Marszalek, C. Schmid, A Spatio-Temporal Descriptor Based on 3D-Gradients, in: BMVC 2008, 2008.

[60] B. Ni, P. Yong, P. Moulin, S. Yan, Multi-level depth and image fusion for human activity detection, IEEE Transactions on System, Man and Cybernetics (B) 43 (2013) 1383–1394.

[61] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32 (2010) 1627–1645.