Computer Vision and Image Understanding 127 (2014) 14–30

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Evaluation of video activity localizations integrating quality and quantity measurements ☆

Christian Wolf [a,b,*], Eric Lombardi [a,d], Julien Mille [a,d], Oya Celiktutan [a,b,e], Mingyuan Jiu [a,b], Emre Dogan [b,f], Gonen Eren [f], Moez Baccouche [a,b], Emmanuel Dellandréa [a,c], Charles-Edmond Bichot [a,c], Christophe Garcia [a,b], Bülent Sankur [e]

[a] Université de Lyon, CNRS, France
[b] INSA-Lyon, LIRIS, UMR CNRS 5205, F-69621, France
[c] Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, Lyon, France
[d] Université Claude Bernard Lyon 1, LIRIS, UMR5205, F-69622, Villeurbanne, France
[e] Boğaziçi University, Dept. of Electrical-Electronics Engineering, Istanbul, Turkey
[f] Galatasaray University, Dept. of Computer Engineering, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 20 April 2013
Accepted 25 June 2014
Available online 9 July 2014

Keywords:
Performance evaluation
Performance metrics
Activity recognition and localization
Competition

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.

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.

☆ This paper has been recommended for acceptance by Anurag Mittal.

* Corresponding author at: Université de Lyon, CNRS, France.
E-mail addresses: christian.wolf@liris.cnrs.fr (C. Wolf), eric.lombardi@liris.cnrs.fr (E. Lombardi), julien.mille@liris.cnrs.fr (J. Mille), oya.celiktutan@boun.edu.tr (O. Celiktutan), mingyuan.jiu@liris.cnrs.fr (M. Jiu), emre.dogan@liris.cnrs.fr (E. Dogan), geren@gsu.edu.tr (G. Eren), moez.baccouche@liris.cnrs.fr (M. Baccouche), emmanuel.dellandrea@liris.cnrs.fr (E. Dellandréa), charles-edmond.bichot@liris.cnrs.fr (C.-E. Bichot), christophe.garcia@liris.cnrs.fr (C. Garcia), bulent.sankur@boun.edu.tr (B. Sankur).

http://dx.doi.org/10.1016/j.cviu.2014.06.014
1077-3142/

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.
- 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). Fig. 1 shows some example frames from this dataset.
- We briefly describe the entry algorithms in the ICPR 2012 HARL competition and we report the evaluation results of the proposed performance metric[2] over these entries, as well as over other baseline algorithms.

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

[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.

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*