Crowdsourcing for Video Event Detection

ROBERT IEPSMA, THEO GEVERS, University of Amsterdam
OANA INEL, LORA AROYO, Vrije Universiteit Amsterdam
ROBERT-JAN SIPS, ZOLTÁN SZLÁVIK, IBM Center for Advanced Studies Benelux

1. INTRODUCTION

With the increased popularity of online video sharing platforms, such as YouTube, and the increase in availability of video capture devices such as smart phones and action cameras, the amount of consumer videos available online is rapidly growing. Many of them are poorly described as it relies mainly on manual effort. The lack of such descriptions creates difficulties for video search systems and ultimately motivates research on automatic content analysis and detection. Tasks, such as, TRECVID Multimedia Event Detection [TRECVID 2016] focus on machine learning approaches for multimedia event detection, to better index video material. However, these tasks typically rely on expert-generated gold standards (GS), which are expensive and time consuming to obtain.

Crowdsourcing proves to be an alternative for gathering gold standards. For example, to gather event-related tags, [Jiang et al. 2011] annotated the Columbia Consumer Video (CCV) dataset through the Amazon Mechanical Turk crowdsourcing platform. Four workers annotated each video, and a voting strategy was used to consolidate the results and filter unreliable labels. This consolidating approach encourages agreement between workers, or bias workers towards predefined labels. We believe that these *agreement-centric approaches are missing out on the semantic diversity of human language* and would not allow for accurate interpretation of events in videos, as they are hardly agreeable due to their intrinsic visual, semantic and syntactic ambiguity. For this, we introduced a *crowdsourcing approach for annotating events in videos that harnesses the disagreement among annotators*, and thus reflects the diversity of expressions and meanings when people refer to events as well as the intrinsic ambiguity of events when they are depicted.

This paper explores how the CrowdTruth [Aroyo and Welty 2014] methodology can be utilized for multimedia event detection. We argue that we can *effectively gather event annotations in videos by harnessing the disagreement among the crowd* and *measure the likelihood or clarity with which a video depicts a given event*. We show how these video event clarity scores can be utilized by a Support Vector Machine (SVM) during training and evaluation.

2. EXPERIMENTAL SETUP

We performed our experiments on a subset of the CCV [Jiang et al. 2011] dataset. We selected a subset of 896 videos (446 videos from the CCV train set and 450 videos from the CCV test set). We selected videos labeled with *parade*, *graduation* or *birthday* as they have the potential to look similar, as well as videos labeled with *soccer* and *basketball* to be able to observe also similarity for sports events. As the videos were too long to use in a crowdsourcing task, we split each video into fragments of maximum 30 seconds. For each video three fragments, evenly distributed across the video, were selected for annotation by the crowd. In total, we annotated 1592 video fragments in a two-step crowdsourcing workflow[1], as described in Section 2.1. We analyze and evaluate the crowdsourcing results using the CrowdTruth methodology. For each task, we save each worker judgment in a vector space representa-

[1]the dataset, methodology and experimental results are accessible here: http://data.crowdtruth.org/events-in-videos/

tion, where the vector is composed of all possible annotations for that task. Furthermore, we compute the unit annotation vector as the sum of all worker annotation vectors for that unit.

2.1 Crowdsourcing Experiments

To annotate the selected video dataset we used CrowdFlower[2]. We run the first task, **Video Labeling**, as an open-ended task, where workers are asked to watch a 30 secs video fragment and provide descriptive labels in a text field. Each worker could annotate videos without being bias towards any predefined event. As preliminary experiments show that the word 'event' appeared ambiguous for the crowd workers, we omitted the word 'event' in the instruction and asked the workers to annotate things that happened in the video.

The results from the open text-field labeling show two types of 'fake' disagreement. We call it 'fake' as workers did not really disagree on what the event was, they simply used diverse expressions to refer to it. For example, workers would use syntactic variations of the same label, or they would use different labels, but refer to the same event. In order to be able to observe the 'true' disagreement, we first performed label standardization, comparison and similarity computation to consider the syntactic variations. To align the use of different labels for the same event we performed a follow-up **Label Clustering** crowdsourcing task.

For the **Label Clustering** task, we first computed the high-level event labels, *i.e.*, the labels with the most agreement among the crowd. Thus, we calculate the average unit-label score for each label: $auls(l) = \frac{\sum_{u \in U_l} uls(u,l)}{|U_l|}$, where U_l is the set of units (i.e., video fragments) in which label l occurs and uls is the *unit-label score*, a cosine similarity score corresponding to the CrowdTruth methodology [Aroyo and Welty 2014] which shows the likelihood of a label l to be expressed in the unit u. The unit-label score is a core CrowdTruth metric that indicates how clear a label is expressed in a video fragment. This formula also allows us to find high-level events with low frequency. We select all labels that have an average unit-label score higher than a certain threshold. In this task, every label gathered in **Video Labeling** is considered a unit. Workers are presented a label and they have to select all the high-level event labels they think it belongs to (multiple choices are possible). The annotation vector is composed of all the high-level events with their frequencies (based on the workers votes). The aim of this crowdsourcing task is to actually transform the **Video Labeling** open-ended task into a closed task where disagreement can be better leveraged.

2.2 Measuring Video Quality Scores

Following the CrowdTruth methodology, we want to see which are the high-level events that are expressed in each video. Thus, we merge the results of the two tasks in a vector space representation. First, we merge the results of the **Video Labeling** task, i.e., we merge all the labels of all fragments belonging to the same video into one vector. Second, we replace each label in this vector with the annotation vector from the **Label Clustering** task that belongs to that label. Third, we compute the annotation vector for each video as the sum of all label-annotation-vectors from the **Label Clustering** task. This vector contains for each label the annotation frequencies depicting each high-level event.

We compute how clear a high-level event is depicted in each video using the video annotation vector. We compute for each video the *video-label-score*, as a CrowdTruth cosine similarity score. This score shows how clear a label is depicted in the video. Since two labels could be similar, the disagreement metric *video-label-score* does not give a correct indication of the label clarity. We account for this by computing the similarity between the labels and incorporating this metric when computing the video-label-score. To compute the video labels we follow the method of [Dumitrache et al. 2015] by setting up

[2]http://crowdflower.com

a threshold for the *video-label-score* and set the final labels used during the training of the classifier for each video to 1 if the *video-label-score* for the corresponding label in that video is higher than the threshold, and 0 otherwise. The *video-label-score* is used during training and testing as a weight.

2.3 Event Classification

We use the same visual and audio features and corresponding feature extraction methods as described and proven useful and complementary for consumer video analysis in [Jiang et al. 2011]. We use a support vector machine (SVM) classifier trained for all classes using the crowdsourced labels. To measure the performance of the classifier, we use the labels above, split into training (CT_{train}) and testing (CT_{test}) sets. We define four sets of videos for training and evaluation. Three different label sets are used for training: CCV: the original CCV labels; CT_{mv}: labels calculated using majority vote over the final merged label vector over the two tasks as described in section 2.2; and $CT@t$: labels calculated using a threshold t on the unit-label scores as described in Section 2.2. In our case, the threshold for the majority vote approach, CT_{mv} is equal to 0.86 as determined by the classifier performance.

3. RESULTS

50,655 labels were gathered from the **Video Labeling** task across all video fragments. 176 (0.3%) were removed as they came from low-quality workers (5/190 workers, 2.6%). The total label set of 50,479 labels contains 5,937 unique labels. After dealing with the 'fake' disagreement (see section 2.1), 5,523 unique labels remained. Most labels occur with a frequency of one (see Figure 1). Further, 192,376 judgments were gathered in the **Label Clustering** task for the 5,523 units. Out of 188 workers, 19 (10.1%) spammers were detected, so we removed their 6714 low-quality judgments.

Each video was annotated by the crowd with an average of 2,897 labels. In Figure 2 we see that the difference in mean *unit-clarity-score* decreases rapidly as the number of workers increases up to 15 workers and stabilizes afterwards. This means that 15 workers are enough to provide the necessary labels. Furthermore, the second crowdsourcing task increased the overall clarity of the labels.

Figure 3 shows the overlap between the labels in the CCV dataset and the crowd labels for each *video-label-score*. The majority of the videos have either labels with high scores ($>$0.8) or low scores ($<$0.2). However, there are also labels with crowd scores lower than 0.8 that also appear in the original CCV labels. Based on the distribution of the labels' scores, the most ambiguous labels with regard to the video seem to have scores between 0.2 and 0.8.

Figure 4 shows the precision, recall and F-score over the three categories, i.e., birthday, parade and graduation. In Figure 4c the $CT@t$ label set shows an increase in F-score as the threshold increases up to a threshold of 0.37. After this threshold it shows only a slight decrease until reaching a threshold of around 0.9 when the F-score drops to 0. This happens due to the small amount of data used. The highest F-score achieved by the $CT@t$ label set is 0.8397, with a threshold of 0.37, which means that allowing more disagreement provides better classification results. We also observe that the $CT@t$ outperforms the majority vote approach, CT_{mv}, between label thresholds in the interval 0.23 - 0.64. Furthermore, with the majority vote approach we are not able to achieve an F-score as good as achieved using the $CT@t$ set. The best performing value for majority vote is only 0.80, compared to 0.8397 as achieved by our disagreement aware approach.

4. CONCLUSION

We introduced a crowdsourcing approach for annotating events in videos that harnesses the disagreement among annotators, reflecting the diversity of expressions and ambiguity when people refer to events. For this we adapted the CrowdTruth metrics to show how clearly each event is depicted in a video. We trained a classifier with crowd labels and showed it achieved good F-score when evaluated with similar data.

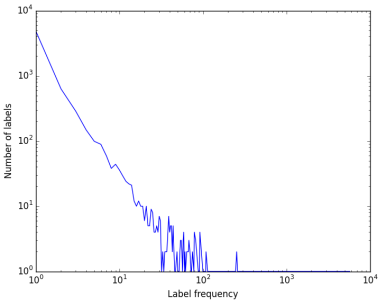


Fig. 1: Label frequency occurrence (log scale)

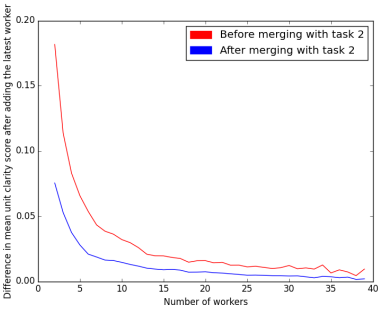


Fig. 2: Difference in the mean unit clarity score after adding a worker for Video Labeling (Task1) task

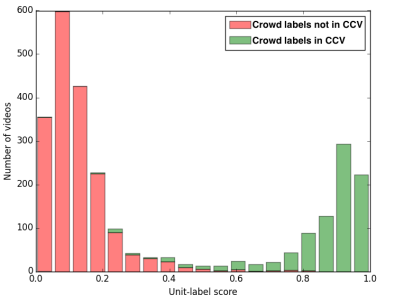
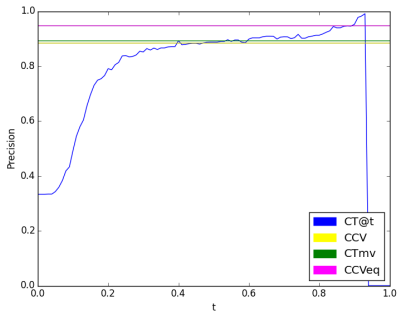
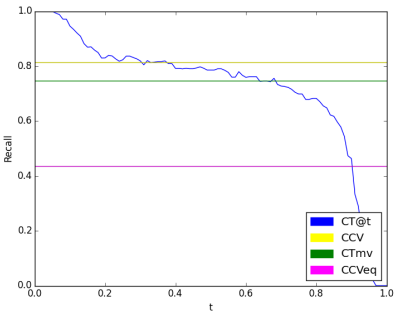


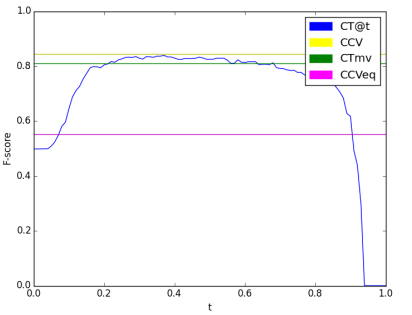
Fig. 3: Frequency of crowd labels that appear or do not appear in the CCV labels at each unit-label score



(a) Precision



(b) Recall



(c) F-score

Fig. 4: Average precision, recall and F-score over the three categories for the support vector machine trained using all three label sets(CT@t, CCV, CTmv) on the CT_{train}-CT_{test} data set

REFERENCES

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Journal of Human Computation* 1 (2014), 31–34.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015. Achieving Expert-Level Annotation Quality with CrowdTruth. (2015).

Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. 2011. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 29.

TRECVID. 2016. Multimedia Event Detection Evaluation Track. http://www.nist.gov/itl/iad/mig/med.cfm. (2016). Accessed: 2016-01-11.