Leveraging Crowdsourced Data for Creating Temporal Segmentation Ground Truths of Subjective Tasks

Matt Burlick Olga Koteoglou Lazaros Karydas
George Kamberov
Stevens Institute of Technology
Hoboken, NJ
{mburlick, okoteogl, lkarydas, gkambero}@stevens.edu

Abstract

We present a new approach to the collection and labeling of ground truth data for annotation of temporal events in ad-hoc videos taken by active operators recording interactions and activities in the field. We present experimental data and related research from experimental psychology which indicate that the conventional methodology based on asking annotators to pick a single instance in time for an event boundary is both unnatural and has several undesirable effects. Our approach is based on allowing the annotators to choose event boundary intervals and modeling each annotators segmentations with mixtures of Gaussians. We use fuzzy measurements to determine an annotators quality and compute a segmentation likelihood function as a Gaussian Mixture of Models (GMMs) over all annotators and boundary intervals. Since the majority of evaluation methods require hard boundaries, we can extract these from the likelihood function as relevant local maxima. We show that given a small set of annotators, this GMM approach provides a more stable ground truth than conventional approaches including majority voting, and demonstrate the application of our approach on two segmentation problems.

1. Introduction

As video recording devices become more portable and affordable there has been an explosion of ad-hoc videos available on the internet. These videos follow very little *production quality* constraints. Namely they often do not have hard cuts between scenes, involve various pans and zooms, and have a wide range of video quality. As a result, performing common video analysis tasks becomes more difficult, as does annotating them.

Historically, annotators have been asked to select a *single frame* to mark an event boundary. Pioneering work in human event perception [3] involved showing several annotators a set of videos and asking them to press a button when they detect an event boundary. However, using this approach, given consecutive events, if two annotators' boundaries differ by even just a single frame their combined effect may be that of over-segmentation.

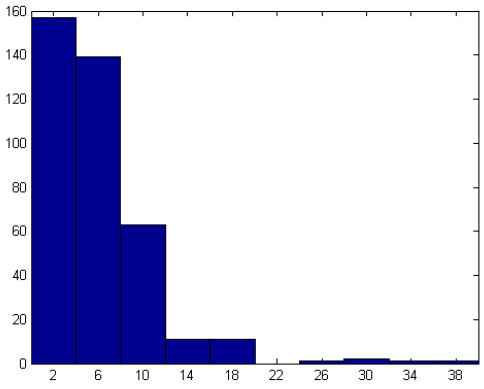


Figure 1: Histogram of segmentation interval widths. x axis represents the interval width in frames and y axis the number of times that width appears.

Zacks, et al. [15, 16], have done extensive work in human event perception. In [15] they provide a hierarchical, recurring, and cyclical model of human event perception. While acknowledging fuzzy temporal boundaries between events and observering that humans process event segmentation at several timescales, there still remains little work on building ground truth segmentation from several crowdsourced annotators that takes into account this issue of fuzzy boundaries.

We argue that especially when dealing with ad-hoc videos, single frame event boundary annotation is unnatural. As evidence, Figure 1 shows the histogram of these boundary interval widths when annotators were given the

978-0-7695-4990-3/13 $26.00 © 2013 IEEE

opportunity to choose *boundary intervals* for the task of *semantic sub-scene segmentation*. Over four annotators on six videos the resulting mean interval width was 7 frames, the median width was 6 frames, and the maximum width was 40 frames. Furthermore the width variance was 12 frames.

This inherent uncertainty in the precise location of the event boundaries can be overcome by performing gaussian smoothing over the boundaries. This is akin to de-noising. Yet we must then choose a value for the Gaussian's σ and by fixing this we loose inherent and independent boundary *confidence*. Instead, if the annotators were to provide *boundary intervals* we could then use those to determine a separate σ value for each boundary. Combined with *annotator quality* we can then model the segmentation likelihood as a Gaussian Mixture of Models (GMM). Finally, from this GMM we can easily create a hard segmentation as relevant local maxima of the likelihood function. An overview of this ground truth extraction pipeline can be seen in Figure 2.

The paper is organized as follows. In Section 2 we first look at related work in the domains of event segmentation and the use of crowdsourced data for establishing ground truths. In Section 3 we look at proposed annotations for the task of temporal segmentation. Section 4 is the heart of our work. It outlines the Ground Truth Extraction Pipeline and all of its components. In Section 5 we explain the experimental setup. Finally in Section 6 we extract ground truths to evaluate the segmentation tasks discussed in Section 5 and show that given boundary intervals and a limited number of annotators, our GMM approach provides more stable ground truths than a conventional majority voting approach.

2. Related Work

As mentioned in the introduction, the traditional approach to event boundary detection is to have several annotators, each of whom *vote* on whether a boundary occurs at each frame. To do this task, Newtson [3] had each annotator click a button upon event boundary detection. In this framework, given $i = 1, ..., N$ annotators, let $L_i(t) \in \{0, 1\}$ be the binary label for frame t chosen by annotators i (0 of there is no boundary, 1 if there is a boundary). A *soft* ground truth boundary can then be computed as:

$$L(t) = \frac{\sum_{i=1}^{N} L_i(t)}{N} \tag{1}$$

From this a *hard* ground truth, L^*, can be obtained by a majority voting scheme:

$$L^* = \{L(t) : L(t) > 0.50\} \tag{2}$$

However, as noted by Zacks, et al.[15], especially in ad-hoc *home movie* style videos, event boundaries can be very *fuzzy*. Therefore it is more desirable to somehow either smooth the boundaries, and/or to use *intervals*.

As leveraging crowdsourced data to provide annotation has become increasingly popular, there has been growing interest and research into questions including "should we consider all annotators as equals?" and "how do we fuse decisions?". Recently, V. Sheng, et al.,[7] discusses the trade-off between quantity and quality of annotators. In order to reduce the number of poor annotators, J. Vuurens, et al.,[11] use qualification tests, trick questions, and time spent on their tasks. Annotators (human or bot) who fail the qualification or the trick question, or who spend too little time on the task are considered unreliable and are dropped from the annotator list. Additionally, annotators who fail to agree with enough other people are also dropped. In the tasks of video ranking, they managed to show that reducing the annotators to mostly *high quality annotators* provides results on par with a true *super-expert* annotator.

V. C. Raykar, et al.,[4] take a similar anti-spamming approach. They attempt to detect spammers in labeling tasks using an empirical Bayesian algorithm that iteratively eliminates the spammers and estimates the consensus labels based only on the good annotators for binary or categorical labels.

Instead of pruning the annotator list, [12, 10] *rank* annotators and use those rankings to weight their contribution to a final decision. [12] uses fuzzy measures (FMs) to model the subjective "worth" of individuals who contribute to crowdsourced data. Their work focuses on providing ratings on things like movies and products, and allows annotators to provide rating intervals. To determine an annotator's worth (or ranking), they look at both the annotator's *specificy* (based on the rating interval width) and *agreement*. They then aggregate these two FMs to create a *meta-measure* and use it in fuzzy integration to determine final ratings.

In [10] the authors explore a set of *annotation scoring functions*. They then score each annotator in one of five ways: number of annotation control points, annotations size, edge detection, Bayesian matting, and object proposal.

The authors of [13] model the task difficulty and annotator competence, expertise, and bias as multidimensional entities. They extract features from images and analyze their parameters among classes and annotators. Their Bayesian approach is heavily reliant on priors and decisions on features to extract.

However, throughout these works only binary, ranking, and rating tasks are addressed.

Our work differs in that we are addressing tasks involving *temporal segmentation* where event boundaries can be *fuzzy* and we model the system as such. As with [4, 12, 11, 10] we are computing annotator quality. Similarly to [12] which allows annotators to specify a *range* or

interval, we argue that an event boundary may not occur at a single moment in time with a likelihood of one, but instead may occur over an interval with varying likelihood. Therefore we model these fuzzy boundaries as Gaussians whose value of σ depends on the interval width. Finally, while common segmentation evaluation metrics require a ground truth with *hard* boundaries, we postpone this decision until the end in order to ensure that all data from all annotators can contribute to the final decision. Similar to [7], this allows us to preserve uncertainty until the final decision.

3. Expert Annotation

For the majority of tasks, in order to perform evaluation a ground truth must be created by an annotator or group of annotators. We call these people *experts* without actually defining what makes someone an expert. Often finding a large number of people who can provide reliable and unbiased ground truth data is difficult and/or expensive. Recently websites, like the Amazon Mechanical Turk (AMT), provide a marketplace where people (know as Requestors) can post tasks and ask *workers* to complete the tasks. While the AMT does allow you to restrict which workers can work on your task according to their credentials (acceptance rate, completion of a moderated tasks, etc.) this does not necessarily guarantee their abilities in completing *your* task, and placing rigid constraints can greatly reduce the number of eligible and willing workers. In the AMT, finding the balance between quality and quantity of workers, as well as monetary incentive, is an art-form unto itself. Nevertheless let us assume that the requestor has figured this out and now wants to create a ground truth from the worker's responses.

In this work we will look at two experiments, which both involve temporal segmentation but have different annotation formats. Let us first look at the simpler case where we ask a user/worker/annotator to indicate each moment in time (frame) that an event boundary occurs, thus segmenting the timeline. While this seemingly goes against our claim that an annotator should specify intervals, we can consider this a subset of full interval annotation where all intervals have a width of zero.

3.1. Boundary Labeling by Annotators

For each video, each annotator $i = 1, ..., N$ looks for event boundaries that would indicate temporal segmentations. Let μ_i^k be the frame at which annotator i detected event boundary number $k = 1, ..., K(i)$, where $K(i)$ is the number of event boundaries detected by annotator i. We ask each annotator to provide their *userID* and a comma separated list of boundary frames:

userID1023: 10, 22, 85, 200

3.2. Interval Labeling by Annotators

There are many event boundaries which cannot be effectively described by a single frame. For the task of determining *semantic sub-scene boundaries*, given the opportunity to choose *intervals*, annotators chose with a mean interval width of 7 frames, a median width of 6 frames, and a maximum width of 40 frames. Furthermore the width variance was 12 frames. The complete histogram can be seen in Figure 1. Choosing intervals for tasks that have ambiguous event transitions, as common in ad-hoc style videos, is significantly more natural than choosing a single frame since it is not clear at exactly which frame the boundary occurs. We also show in Section 6.1 that given limited annotators who provide intervals, we can use this information to create ground truth data that is more stable than simply using a majority voting approach based on single-frame boundaries.

Let us redefine our labeling problem. For each video, each annotator $i = 1, ..., N$ looks for an event boundary interval that would indicate a temporal segmentation. Let s_i^k and e_i^k be the start and end frame of the *k-th* interval, respectively. Each annotator can then provide a single line stating their *userID* and a comma separated list of $[s_i^k, e_i^k]$ intervals. For example:

userID1023: [8, 12], [20, 23], [75, 90], [199, 201]

4. Ground Truth Discovery Pipeline

Once we have labeled data from our annotators we then need to determine a ground truth using their data. Since not all workers are created equal (in fact some may even be malicious!) and the fact that mislabeling an event boundary by a single frame can have a negative effect, we first model a soft ground truth, or event segmentation likelihood function, as a Gaussian Mixture of Models (GMMs). Each event boundary interval for each annotator provides a Gaussian, and the Gaussians are mixed according to their *annotator's quality*. This quality in determined using the Fuzzy Measurement of Agreement as demonstrated in [12]. Once we have established an event segmentation likelihood function, we transform it into a hard ground truth segmentation by finding local maxima and keeping only the local maxima which exhibit at least 50% likelihood relative to the maximum likelihood.

Figure 2 shows the ground truth discovery pipeline. We will now examine each of these steps in detail.

4.1. Determining Annotator Quality

The Fuzzy Measurement of Agreement measures *how often an annotator agrees with the group*[12]. To determine this, for each annotator we observe how much of their segments agree with the segmenents of others. Let $\chi_i(t)$ be the *segment label* assigned by annotator i at time t. This segment label can either be the segment number that the frame

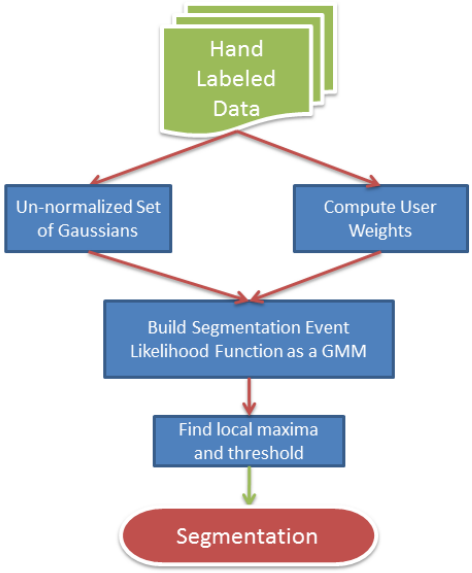


Figure 2: The Ground Truth Discovery Pipeline

falls within and/or can contain some *semantic information* about the segment (for example, the relevant characters in the video). Regardless, given T frames, we can compute how often an annotator i agrees with all other annotators as:

$$a_i = \sum_{j=1}^{N} \sum_{t=1}^{T} (\chi_i(t) == \chi_j(t)) \tag{3}$$

We can then normalize this to provide our annotator qualities:

$$\alpha_i = \frac{a_i}{\sum_{j=1}^{N} a_j} \tag{4}$$

If we have $p = 1, ..., P$ event boundaries accumulated over all N annotators, we can then let $W(p) = \alpha_i$ be the annotation quality (confidence) associated with event boundary p which was generated by annotator i. Note that we can also opt to set $W(p) = \frac{1}{N}$ in order to set uniform quality across all annotators.

In our experiments with semantic sub-scene segmentation, which had four annotators, we found the values of α to be in the range of 0.11 to 0.30.

4.2. Building a Temporal Event Likelihood Function

We model our event segmentation likelihood function as a Gaussian Mixture Model where we have one Gaussian per event boundary and weight the Gaussians according to the annotator's quality, using Equations 3 and 4.

If we fix the value of σ for each Gaussian (as would be the case if the annotators only provide a single frame), the effect of using Gaussians to model events boundaries is akin to Gaussian smoothing with a δ function centered at each boundary location with its height based on its annotator quality. However if an annotator provides *intervals* we can use these to compute a σ value for each Gaussian.

Given annotations in the interval format, for a given annotator i's k^{th} boundary, let $\mu_i^k = \frac{e_i^k + s_i^k}{2}$ and $w_i^k = e_i^k - s_i^k$ be the boundary center and width, respectively. Then using basic statistics and the *quantile function* we can determine the value of σ_i^k given the interval width w_i^k and the desired capture probability p. Let us define the *inverse probit function* $\phi^{-1}(p)$ as

$$\phi^{-1}(p) = \sqrt{2}\,\mathrm{erf}^{-1}(2p-1), p \in (0,1) \tag{5}$$

where erf is the error function. If we denote the *quantile* $z_p = \phi^{-1}(p)$ we can then say that a normal random variable X will exceed $\mu + z_p\sigma_i^k = \mu + \frac{w_i^k}{2}$ with probability $1-p$. Therefore, given capture probability p, we can compute z_p and determine σ_i^k as:

$$\sigma_i^k = \frac{w_i^k}{2 * z_p} \tag{6}$$

In our work we chose $p = 0.90$.

Let $\mathcal{U}$ and Σ be the **collection** of all event boundary centers and their respective standard deviations across all annotators. We can then compute the event segmentation likelihood function as:

$$L(t) \propto \frac{1}{\|\mathcal{U}\|} \sum_{i=1}^{\|\mathcal{U}\|} W(i) \frac{1}{\Sigma(i)\sqrt{2\pi}} e^{-\frac{(t-\mathcal{U}(i))^2}{2\Sigma(i)}} \tag{7}$$

where $W(i)$ is the weight of event boundary i based on its annotator's quality as computed in Section 4.1.

Figures 3 and 4 show the event segmentation likelihood graph for a single annotator for the task of semantic video sub-scene detection (see Section 5.1). In the first figure the annotator specifies the event boundary interval (and thereby its value of σ_i^k), whereas in the second figure they only provide a single frame boundary and a fixed value of $\sigma_i^k = 0.6$ is used. Note that this value of σ_i^k could also be a parameter that can be tuned based on event boundary ambiguity.

Figure 5 shows the final event segmentation likelihood graph based on four annotators who provided intervals and whose quality were computed as in Section 4.1.

4.3. A Hard Segmentation Ground Truth from Event Segmentation Likelihood Function

Having a final hard segmentation ground truth allows for application of many evaluation metrics and measurements.

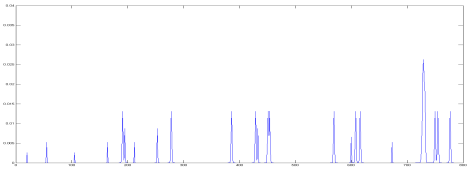


Figure 3: Single annotator's segmentation likelihood function in frames [0 800] based on event boundary *intervals*

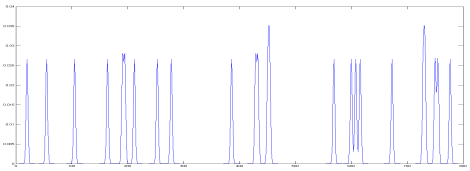


Figure 4: Single annotator's segmentation likelihood function in frames [0 800] based on event boundary *frames* ($\sigma = 0.60$)

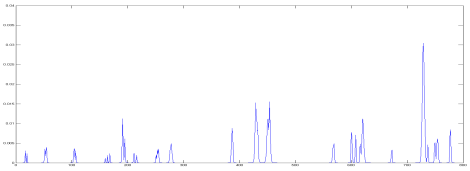


Figure 5: Multi-person event segmentation likelihood function in frames [0 800] based on annotator event boundary *intervals*

Therefore it is imperative that we are able to extract a hard segmentation from our likelihood function. To do this we first find the local maxima in the likelihood function. These represent *potential* boundaries. However we are only interested in *highly likely* boundaries, in particular boundaries that have at least 50% likelihood relative to the most likely boundary (though hypothetically this could be a parameter that a user could tweak in order to establish ground-truth sensitivity). Therefore we filter out all local maxima that are below this level of likelihood, leaving us with a final hard ground truth.

Figure 6 shows the likelihood and hard segmentation graphs based on four annotators who provided intervals for a video with 1194 frames. The top graph is the likelihood graph as created in Section 4.2 and the bottom graph is the hard segmentation extracted from the likelihood function as described in this section.

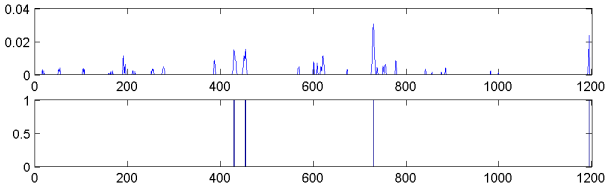


Figure 6: Multi-person segmentation based on annotator event boundary *intervals* with weighting by annotator quality

5. Experimental Setup

In this work we looked at two problems where we need an event segmentation ground truth in order to evaluate the segmentation quality of different algorithms.

5.1. Semantic Sub-Scene Segmentation

The first task is that of segmenting a video into *semantic sub-scenes*. We define a semantic sub-scene as a sequence of frames that all have the same *principal characters*. When either new principal characters enter the frame, or one or more leave it, the current semantic sub-scene terminates and a new one is spawned.

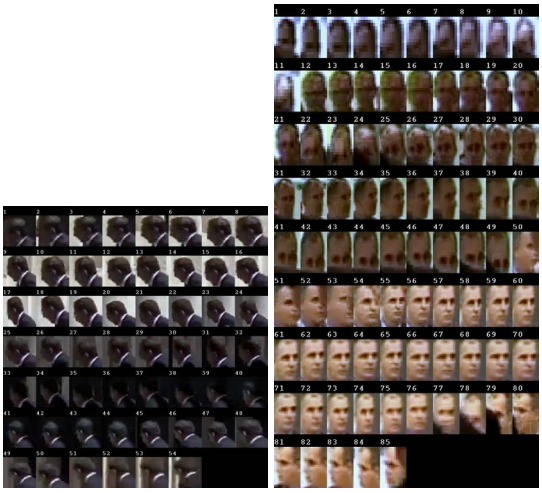
The collection of videos we applied semantic sub-scene segmentation to, consists of ad-hoc videos with hand-held camera and low resolution. Two frames of an example video can be seen in Figure 7.

(a) (b)

Figure 7: Example frames in Annarella video

5.2. Object Track Segmentation

Even state-of-the-art object trackers are highly flawed. When faced with appearance or illumination changes as well as partial or entire occlusion, trackers exhibit errors including false positives, false negatives, association, and drift. Therefore it is often quite important to identify subsequences of object tracks that are stable. This is the task of object track segmentation. Figure 8 shows examples of tracked objects (people).



(a) (b)

Figure 8: Examples of tracked objects

5.3. Evaluation

Let us define the ground truth and automatic segmentations as:

$$X = \{x_1, ..., x_{N_X}\} \tag{8}$$

$$Y = \{y_1, ..., y_{N_Y}\} \tag{9}$$

respectively. We can then use the *purity*[9], π, which is a performance metric commonly applied in segmentation problems, to evaluate the quality of our segmentations. Given the ground truth and automatic segmentations the purity π is defined as:

$$\pi = \left(\sum_{i=1}^{N_X} \frac{\tau(x_i)}{T} \sum_{j=1}^{N_Y} \frac{\tau^2(x_i, y_j)}{\tau^2(x_i)} \right) \cdot \left(\sum_{j=1}^{N_Y} \frac{\tau(y_j)}{T} \sum_{i=1}^{N_X} \frac{\tau^2(x_i, y_j)}{\tau^2(y_j)} \right) \tag{10}$$

where $\tau(x_i, y_j)$ is the length of the overlap between the time interval corresponding to segments x_i and y_j, $\tau(x_i)$ is the length of the time interval corresponding to segment x_i, T is the total length of all the segmentations. In each parenthesis, the first term is the fraction of the current evaluated story, and the second term indicates how much a given story is split in to smaller stories. Purity values are in $[0, 1]$. The closer it is to 1 means greater similarity between the automatic segmentation and the ground truth.

6. Results

In Section 6.1 we analyze ground truth *stability* for both our GMM-based and a majority voting approach. We define a stable method as one that results in similar ground truths independent of the set of labelers. Using this stability measurement, we do in fact show that given only a few annotators who provide event boundary intervals, our GMM-based method is more stable than a conventional majority voting approach.

We then look at the application of our temporal ground truths for evaluating segmentation algorithms for the two tasks mentioned in Section 5. Note that since the purity metric measures similarity between two segmentations, we only compare against our GMM-based method. We do not compare the resulting purity values against a ground truth built using majority voting since purity only measure how well two segmentation *match*. As a result, a higher purity measurement using different ground truths does not necessarily imply that the ground truth that provided the higher purity measurement is more accurate. It only implies that it matches the automatic segmentation better.

6.1. Analysis on Ground Truth Stability

In this work we have made the claim that providing boundary *intervals* for event segmentation allows for more robustness to annotation noise. For largely subjective tasks, like that of annotating events in ad-hoc videos, a ground truth can be determined by leveraging crowdsourced annotations. We argue that we can assess the quality of our ground truth by observing how *stable* the generating process is. We measure this stability by building several ground truths using different annotators and determining how similar they are to one another.

Given $A = \{1, ..., N\}$ annotators, let us choose a subset of annotators at random, $B \subset A$. From this subset of annotators we can build a ground truth. We can then repeat this process M times to generate M ground truths. Let $L^*_{(gmm,m)}$ be the m^{th} hard ground truth segmentation using our approach in Section 4 and $L^*_{(mv,m)}$ be the associated ground truth using Majority Voting as in Equations 1 and 2. Then between any two ground truths, we can compute a purity measure $\pi\left(L^*_{(\cdot,i)}, L^*_{(\cdot,j)}\right)$ and define the set of pairwise purities as:

$$\pi_\cdot = \left\{ \pi\left(L^*_{(\cdot,i)}, L^*_{(\cdot,j)}\right) | i = 1, ..., M, j = 1, ..., M \right\} \tag{11}$$

Table 1 shows the statistics for π_{gmm} and π_{mv}. The ground truths were generated using the annotations from semantic sub-scene segmentation experiment (see Section 6.2). From this we can gleam that our GMM based approach is more stable than a Majority Voting approach due to the higher average purity and lower standard deviation of purity, as computed between ground truths built on randomly selected annotators.

	GMM Based GT π_{gmm}	Majority Voting GT π_{mv}
Min	0.430	0.168
Max	1.000	1.000
Median	0.944	0.660
Mean	0.859	0.711
Std	0.173	0.242

Table 1: Statistics on ground truth set purities, π_{gmm} and π_{mv}, as computed in Equation 11.

Method	Sub-Scenes	MDL	ShoryShed
Annarella	0.580	0.815	0.506
Annarella2	0.194	0.194	0.560
David_Blaine	0.118	0.118	0.179
Quartli	0.151	0.431	0.264
Venizelos_clip1	0.663	1.0	0.990
Venizelos_clip2	0.371	0.912	0.989
Overall	0.346	0.624	0.581

Table 2: Purity Measurements for three semantic segmentations against our ground truth segmentation. Larger values are better.

6.2. Application: Semantic Sub-Scene Segmentation

Here we asked four annotators to provide event boundary *intervals* for six different ad-hoc videos. For each video, we determined annotator quality, created an event segmentation likelihood function, and finally produced a hard ground truth segmentation. We then segmented each video into *semantic sub-scenes* based on character participation using a proprietary algorithm. Next we used two methods to discover *semantic scenes*: Minimal Description Length[5], and Story Shed[14]. Finally, using the hard ground truth and the purity measurement, we compared the quality of these different segmentations. The results can be seen in Table 2.

6.3. Application: Object Track Segmentation

Here we made available on the Amazon Mechanical Turk 36 different object tracks. The number of annotators who attempted to segment each track ranged from 4 to 57. The annotators were only asked to specify event boundaries in the format outlined in Section 3.1. Since the annotators only provided single frame boundaries, we fixed the interval width to be three. This design choice was made by subjective observation of the data and could be a parameter varied according to desired uncertainty. From the annotations we then generated a hard ground truth segmentation.

Against this ground truth we compared five different segmentation algorithms. These include global agglomerative clustering methods, like *mean shift*[1], the *L-Method*[6], the *Gap Statistic*[8], and the method used in [14] to determine the number of main characters, which we call the *jump method*. We also look at a local method where we segment a track whenever adjacent (temporally) features have a greater value of *disjoint information* than they do *mutual information*. For each of the methods we looked at several different features (see Table 3).

Algorithm	Feature Type	Purity
IT	Sum of RGB Histograms	0.400
Meanshift	RGB Histogram	0.265
L-Method	Hue Pixels	0.196
Jump-Method	Indexed RGB Histogram	0.142
Gap Statistic	Sum of RGB Histograms	0.446
Shot Detect		0.457

Table 3: Object Track Segmentation Purity

Finally we also use a shot boundary detection algorithm [2] on a video generated using the tracked object crops, resized to 30x30.

To evaluate the quality of the segmentations we use the purity measurement 10 .

Figure 9 shows the ground truth segmentation of the object in Figure 8b against the shot detect[2] algorithm's segmentation and Table 3 shows the purity of each method against the computed ground truth.

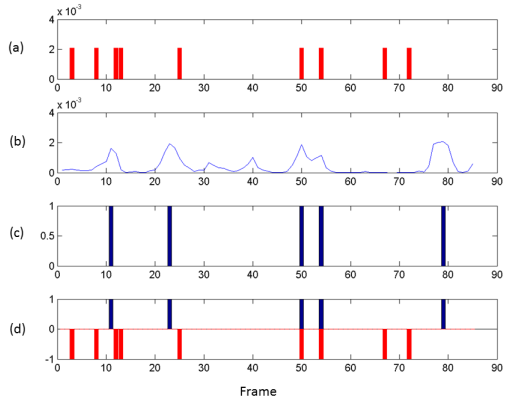


Figure 9: Segmentation for object track show in Figure 8b. (a) is the segmentation according to the shot detect [2] algorithm. (b) is the ground truth segmentation likelihood function. (c) is the ground truth hard segmentation. (d) shows (a) and (c) together.

7. Conclusions

In this work we presented a novel way to create ground truth data for temporal event segmentation using crowd-

sourced annotations. We demonstrated its application to two different segmentation problems and showed that by using event boundary intervals, we are able to produce ground truths that are more stable than with majority voting. While the applications we demonstrated are related to computer vision and human perception, the concepts of this work are meant to be applicable to various research fields and problems, especially those that have the need for segmentation of temporal data in the face of ambiguous, fuzzy, and subjective boundaries.

Moving forward we aim to do further analysis on our ground truth segmentations. In particular to observe the effects of annotator quality and quantity on ground truth stability. Much like done in the work by Zacks, et al., we are also interested in using our approach to analyze segmentation tasks in a variety of disciplines.

References

[1] Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995. 7

[2] J. Mathe. Shotdetect. http://johmathe.name/shotdetect.html. 7

[3] D. Newtson. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28(1):28–38, Oct. 1973. 1, 2

[4] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13:491–518, Mar. 2012. 2

[5] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465 – 471, 1978. 7

[6] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576 – 584, nov. 2004. 7

[7] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 614–622, New York, NY, USA, 2008. ACM. 2, 3

[8] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Royal Statistical Society*, 2001. 7

[9] A. Vinciarelli and S. Favre. Broadcast news story segmentation using social network analysis and hidden markov models. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 261–264. ACM, 2007. 6

[10] S. Vittayakorn and J. Hays. Quality assessment for crowdsourced object annotations. In *Proc. BMVC*, pages 109.1–109.11, 2011. 2

[11] J. Vuurens and A. de Vries. Obtaining high-quality relevance judgments using crowdsourcing. *Internet Computing, IEEE*, 16(5):20–27, 2012. 2

[12] C. Wagner and D. Anderson. Extracting meta-measures from data for fuzzy aggregation of crowd sourced information. In *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, pages 1 –8, june 2012. 2, 3

[13] P. Welinder, S. Branson, S. Belongie, and P. Perona. The Multidimensional Wisdom of Crowds. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2424–2432. 2010. 2

[14] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Rolenet: Movie analysis from the perspective of social networks. *Multimedia, IEEE Transactions on*, 11(2):256–271, 2009. 7

[15] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds. Event perception: A mind/brain perspective. *Psychological Bulletin*, 133:273–293, 2007. 1, 2

[16] J. M. Zacks and B. Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127:3, 2001. 1