© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1722842

Discontinuity is not a cut: automatic shot boundary detection and the measurement of film editing

MD RIAZ UDDIN[1], SAMEER AHMED[2]
[1, 2]*Department of Film and Television, Jagannath University, Dhaka, Bangladesh*

Abstract- ***Film studies increasingly computes average shot length from automatically detected shot boundaries. Detectors do not identify shots. They identify discontinuities in the visual signal, which coincide with shots only approximately. Because that divergence is structured by the footage rather than random, it biases shot-length statistics rather than adding noise. Using the ClipShots ground truth (5,138 shots across 92 videos) with detector performance figures published for that corpus, we simulate how detection error propagates into the statistics the field reports. In these simulations average shot length is systematically deflated, by 15.4% and 10.6% for the two stronger detector profiles, because false boundaries outnumber missed ones. Video is made to appear as though it cuts faster than it does. The bias does not track detector quality, and it is confounded with the object of study. Since gradual transitions are missed far more often than cuts, a film's bias varies almost perfectly with the proportion of its transitions that are gradual, itself a stylistic property. Films that cut hard and films that dissolve are distorted in opposite directions by the feature distinguishing them. Differences below about 5% prove unrecoverable under every profile tested. Researchers comparing moving images on shot length cannot assume the size of their measurement error, nor remove it by choosing a detector with a better F-score.***

Index Terms- ***average shot length, cinemetrics, film editing, measurement error, shot boundary detection***

I. INTRODUCTION

Cuts are everywhere in film and largely invisible. A feature contains a thousand or more of them, and continuity editing exists precisely to make them pass unnoticed: viewers asked to watch for cuts miss between a tenth and a half of them, depending on the type of cut, as reported in [7]. The cut is nonetheless the unit on which the quantitative study of film style rests. Counting the cuts in a film and dividing its running time by the number of resulting segments gives the average shot length, the oldest and most widely used measure in the field and one intended to capture something real about pace. A film averaging two seconds a shot is doing something different from one averaging ten.

Obtaining that number used to be laborious. Someone had to sit with a film and a stopwatch, or later a keyboard, marking every cut across two hours of running time. Because the work was slow it was rare, and because it was rare, arguments about film style remained largely arguments. Software then arrived that could do the marking automatically, in minutes, on any film supplied to it. The cost of the measurement collapsed, and the number of studies using it rose accordingly. A recent tool paper describes a pipeline that segments films with a neural detector and reports average shot length from the result [12]. A dataset published in 2026 distributes shot boundaries and duration statistics for thirty Kurosawa films, produced the same way [29]. Neither is unusual and neither is careless by the standards of the field. Both report shot-length statistics as measurements.

The difficulty that concerns this paper lies in what that software does. It is not looking for cuts. It is looking for moments when the picture changes sharply from one frame to the next. Most of the time a sharp change means a cut, which is why the approach works at all. But not always. A flash of lightning changes the picture without any cut having occurred. So does an explosion, a camera flare, or a hand passing in front of the lens. Conversely, a cut between two similar shots (the same framing, the same light, the back-and-forth of an ordinary conversation scene) barely changes the picture at all, and can slip past unnoticed. The machine and the viewer are not counting the same events.

How often they disagree is not a secret. Computer scientists have measured it carefully for twenty-five years and publish the figures, sometimes in the same documents that distribute the data. What has attracted far less attention is what those disagreements do to the numbers published on the other side. That is a simple question with an arithmetic core. If the software misses a cut, two shots are counted as one, and the film appears to have longer shots than it does. If the software imagines a cut, one shot is counted as two,

and the film appears to have shorter shots. Average shot length is running time divided by the number of shots, so miscounting shots moves it directly.

We expected the misses to dominate, since gradual transitions such as dissolves are known to be the hardest cases. We were wrong. Using error rates published for the material we studied, false cuts outnumber missed ones, and average shot length comes out too low, by 15% for the strongest detector we examined. The material is made to look as though it cuts faster than it does.

The direction is awkward, because it runs the same way as a claim the field has been making. Cutting rates are widely reported to have accelerated over the last century, and short-form platform video is now commonly described as cut at only a few seconds a shot. A tool that manufactures fast cutting is being used to demonstrate fast cutting. We are not saying those findings are wrong. The historical acceleration in particular is far too large to be an artefact. We are saying that the measurement error points the same way as the hypothesis, which is the least comfortable place for it to sit.

A third result proved less expected than either. The size of the error is not the same for every film. It depends almost entirely on how much a film uses dissolves and fades rather than straight cuts, and that is a stylistic choice, not a technical property. Films that cut hard and films that dissolve are distorted in opposite directions by the very feature that distinguishes them. The error, in other words, is not noise sitting on top of the thing being measured. It is entangled with it.

The rest of this paper sets that argument out properly. Section 2 asks what a shot boundary detector is actually identifying and why it is not quite a shot. Section 3 places the problem between two research literatures that have not spoken to each other, and borrows a method from a third that has. Sections 4 and 5 give the method and results, and Section 6 works through what follows for claims the field has already made.

II. WHAT A DETECTOR THINKS A SHOT IS

The argument to this point has been about counting. It is really about a concept, and the rest of this section makes that case in the terms film studies uses for such questions: what the shot is taken to be, who decides, and what changes when the deciding is handed to a machine. Readers who want the measurements can take Section V on its own.

Cinemetrics rests on a unit it rarely examines. Salt's founding paper set out to replace what he called a matter of loose assertion rather than demonstration with something measurable, and in choosing what to measure he named two criteria: the variables most directly under the director's control, and, in his own qualification, those easiest to quantify [21]. The two criteria are not the same, and the history of the field since has largely been the story of the second quietly constraining the first. Shot length prevailed because it was tractable.

What counts as tractable has since changed hands. For most of the field's history, shot boundaries were determined by a person watching a film and pressing a key: slow, tiring, inconsistent, but at least an operation in which the unit being counted is the one film theory recognises. A human annotator counting shots is applying a concept. Ease of quantification now means ease for a detector, which is a different constraint altogether, and one that no film theory has had any part in setting.

A detector is not applying that concept. It is applying a different one that mostly coincides with it. Shot boundary detection algorithms, in nearly every formulation from the histogram-difference methods of the 1990s to the convolutional networks of the last decade, identify discontinuities in the visual signal. They measure where the image changes abruptly. The premise is that images change abruptly when and only when the camera changes, and the premise is a good approximation, which is why the technology works at all. But it is an approximation, and its failures are not random.

The cases that fall into the gap are instructive. A dissolve is, for a film scholar, a single articulation joining two shots. For a detector it is a region tens of frames long within which no single frame is the boundary. Placing a point somewhere in that region is a convention rather than a discovery, and the hand-logging tradition has its own: Redfern notes that in cutting-rate work the moment of transition is taken as the midpoint of a dissolve [20]. The difficulty is not that conventions are used. It is that the human convention is stated and the detector's is not, and there

is no guarantee the two agree. The cases named in the introduction, lightning and lens flares on one side and matched shot-reverse-shot on the other, are instances of the same problem. So is the whip pan used as a transition, which is genuinely ambiguous in film practice and about which the detector must nonetheless decide.

The point is not that detectors are unreliable. It is that a detector and a film scholar are not counting the same thing. The detector segments the film into stretches of visual continuity. The scholar wants it segmented into shots. These two partitions agree most of the time, and the published precision and recall figures are, read properly, a measurement of how often they disagree. What the computer vision literature reports as detector error is, from the standpoint of film studies, the size of a conceptual gap.

Reading the error rates this way changes what one expects of them. If the mismatch were random noise, it would be symmetric, and its effect on shot-length statistics would tend to cancel. It is not random. It is structured by properties of the footage (motion, occlusion, lighting instability), and structured error biases a statistic rather than merely widening its spread. That is the claim this paper tests, and it holds.

Which way the bias runs is a separate question, and there our expectation was wrong. We reasoned from the best-known weakness of these tools: gradual transitions are the hardest cases, dissolves defeat detectors that hard cuts do not, so misses should dominate and shot length should come out too long. What that reasoning overlooked is that the same turbulence which hides real transitions also manufactures apparent ones. Once a detector's precision is weaker than its recall, the ordinary situation on difficult material, false boundaries outnumber missed ones and the bias inverts. The mechanism we had identified was real. We had followed only one of its two consequences.

There is a second reason to state the problem in these terms. Cinemetrics has understood itself, at least since Tsivian's platform made large-scale comparison possible, as a shift from interpretation toward measurement, from arguing about what a film means to establishing what a film does [4]. The appeal of the shift is that measurements are supposed to be stable in a way that interpretations are not. But a measurement is only as stable as the operation that produces it, and when the operation changes, from a person with a keyboard to a network trained on YouTube video, the quantity being measured changes with it, whether or not the name stays the same. Two average shot lengths for the same film, one hand-logged and one detected, are not two estimates of one number. They are measurements of two related but distinct properties.

Digital humanities has a standing version of this worry, though it is not the dominant position. Bakels and colleagues set out three challenges for computational analysis of audiovisual media, the first of which is recasting film-analytical vocabulary into a machine-readable ontology [2]. That is precisely the operation at issue here, performed silently: the detector embodies an ontology of the shot, it was never written down, and film studies has been using its output as though the translation were transparent. Chávez Heras makes the stronger methodological case, following Agre, that computational film scholarship should proceed as critical technical practice, in which the construction of the instrument forms part of the argument rather than a preliminary to it [3]. This paper is a small exercise in that mode: we take an instrument the field uses without examining it and ask what it does to the claims made with it.

It should be said that the prevailing assessment is more sanguine. Pustu-Iren and colleagues survey automated visual analysis for film studies, compare machine against human performance on several recognition tasks, and conclude that the results make a strong case for adopting these methods, noting that the approaches remain prone to errors of varying degree [15]. We do not dispute the case for adoption, since the tools make work possible that would not otherwise be done. Our argument concerns what that concession to error implies once the erroneous output is aggregated into a statistic and compared across films, which is a question the survey does not take up.

The field has been here before, on a smaller scale, and handled it well. Cutting and colleagues reported that shot lengths were shaped by act structure, with peaks at the quarter, half and three-quarter points of films [5]. Salt tried to replicate it, and suggested the peaks were an artefact of the procedure used to normalise films of different lengths onto a common scale. The authors reanalysed with three interpolation methods and a different binning scheme, confirmed that the

peaks vanished, and stated that there is no support in the shot lengths for the claim that the complication and development sections shape them. Other findings in the original paper survived. The retraction was partial, and precisely bounded. They closed by thanking Salt for pointing out the error [8]. That exchange is a model. It also demonstrates that measurement artefacts in this field are capable of producing findings that look like style. The difference is that the artefact in that case lived in a step the authors had written themselves and could inspect. The artefact examined here lives inside a detector most users did not write and cannot easily inspect, and it is applied at the very first step, before any analysis begins.

A final observation about the founding text bears on all of this. Having settled on shot duration and shot scale, Salt observed that his analyses could be extended, and that an obviously important quantity would be the strength of cut, or as he immediately corrected himself, the nature of the shot transition from each shot to the next [21]. He set the problem aside as difficult and proceeded with what he could count. Fifty years later the character of the transition turns out not to be an optional extension of the measure he adopted but the thing that governs whether it can be trusted at all, which is what Section V.D shows.

III. A GAP BETWEEN TWO LITERATURES, AND A METHOD FROM A THIRD

The problem sits between two bodies of work that have had little to say to each other.

On one side is the evaluation of shot boundary detection, which is thorough. TRECVid ran a dedicated shot boundary task for seven consecutive years, drawing fifty-seven research groups onto common data and common scoring [24]. Lienhart's comparative study, now more than twenty-five years old, had already established the pattern that has held since: hard cuts and fades are detected reliably, dissolves are not, and the false hit rate for dissolves in the systems then available ranged from 50% to more than 400% [14]. Hanjalic set the problem out formally shortly after [10]. A recent review consolidates two decades of subsequent work and reaches familiar conclusions about illumination change and camera motion as the dominant sources of failure [11]. None of this literature asks what its error rates do to anything computed downstream, because nothing is computed downstream: shot boundaries are a preprocessing step on the way to retrieval or summarisation, and the boundaries are the output of interest.

On the other side is the statistical study of film style, where the fragility of shot-length measures has been argued at length, though a different fragility. Redfern has shown that shot-length distributions are not adequately modelled by the log-normal, examining 134 Hollywood films and finding the assumption unsupported in 125 of them [18]. He has argued for dominance statistics over comparisons of central tendency [17], and for thinking about shot lengths distributionally rather than through single summary values [19]. The concern throughout is with what one does to a set of shot lengths after obtaining them: which estimator, which comparison, which assumption. The shot lengths themselves are taken as given.

So the two literatures divide the problem between them and leave a gap in the middle. Computer vision knows how often the boundaries are wrong. Cinemetrics knows how easily shot-length statistics mislead. To our knowledge neither has examined what happens when wrong boundaries feed statistics that mislead easily.

The method for answering it comes from a third place. Communication research has spent the past decade on structurally the same problem: a classifier assigns labels, the labels are aggregated, and the aggregate becomes a variable in a subsequent analysis. TeBlunthuis and colleagues show that classifier misclassification biases downstream regression estimates, document through systematic review that the field has largely ignored this, and supply correction methods [27]. Bachl and Scharkow had earlier demonstrated matrix back-calculation for the same purpose in content analysis [1]. Wang and colleagues formalise the general case of inference on machine-predicted outcomes [28]. Qiao and Huang address the specific structure in which individual predicted labels are aggregated into a group-level variable [16], which is what average shot length is: a per-film aggregate over per-boundary predictions.

What we import from this work needs stating precisely, because it is less than it might appear. TeBlunthuis and colleagues are concerned with misclassified variables entering a regression as independent or dependent variables, and the

correction methods they compare (regression calibration, multiple imputation, and the maximum likelihood adjustment they propose) are estimators for recovering unbiased regression coefficients. Average shot length is not a regression coefficient. It is an aggregate computed over per-boundary predictions, which is structurally closer to the case Qiao and Huang treat.

We therefore borrow two elements and not a third. The first is the diagnosis: that classifier error propagates into downstream estimates, that it does so systematically rather than as noise, and that a field can adopt classifiers enthusiastically while ignoring this for years, which TeBlunthuis and colleagues establish for communication research through a systematic review of forty-eight empirical studies. The second is the design of the demonstration: Monte Carlo simulation over known error rates, which is how all of this work proceeds. What we do not borrow is the correction machinery itself, for a reason given in Section VII.

IV. DATA AND METHOD

A. Corpus

We use the test split of ClipShots [26], a shot boundary dataset assembled from YouTube and Weibo and annotated for both abrupt and gradual transitions. ClipShots was chosen for three reasons. Its annotations are distributed openly and separately from the video, so the analysis requires a few megabytes rather than a few hundred gigabytes. Its transition annotations record the frame span of each transition, which allows abrupt and gradual transitions to be distinguished directly from the data rather than assumed. And detector performance figures for this exact corpus are published alongside it, so the error rates we inject were measured on the material we inject them into.

Each record gives a list of transitions as start and end frame pairs, together with the video's frame count. We treat a transition whose span is one frame or less as an abrupt cut and anything longer as gradual, place the boundary at the midpoint of the span, and derive shot durations as the intervals between consecutive boundaries. Durations are in frames, since the dataset supplies no frame rate. This does not affect the results, which are reported as proportional bias and are therefore invariant to frame rate.

The test split comprises 500 videos with 5,876 cut transitions and 2,422 gradual transitions, a gradual share of 29.2% [26]. Every video was double annotated for quality assurance, which is more than most benchmarks of this kind offer.

Average shot length is not meaningful for a video with a handful of shots, so we restrict to videos with at least thirty. This leaves 92 videos and 5,138 shots, of which 3,738 are followed by an abrupt cut and 1,308 by a gradual transition. The remaining 92 are each video's final shot, which is followed by no transition. The gradual share is 25.9%. Relaxing the threshold to ten shots yields 194 videos and 6,914 shots at 29.8% gradual, close to the figure for the split as a whole. Results are stable across both thresholds.

One split of ClipShots must be avoided. According to the repository documentation, though the point is not made in the paper, the `only_gradual` set annotates gradual transitions alone, as a supplement compensating for their scarcity in the training data, and its abrupt cuts are unlabelled. Shot lists derived from it are not shot lists, and any statistic computed from them is meaningless. We mention it because the file is distributed alongside the others and its name does not obviously warn against the use.

B. Error profiles

The ClipShots documentation reports precision and recall for three detectors, separately for cut and gradual transitions. Those figures use the standard TRECVid matching criterion: a predicted boundary counts as a hit if it overlaps the ground-truth transition by at least one frame [26]. Given one-to-one matching under that criterion, recall is the proportion of true transitions found and precision the proportion of predictions that are real, so we convert directly: miss rate is one minus recall, and the ratio of false positives to true positives is (1 − precision) / precision.

Table 1. Published precision and recall for three detectors on the ClipShots test split [26], with the miss and false-positive rates derived from them. DSM is deep structured models, and the two deepSBD variants are named by backbone.

Measure	DSM	ResNet-18	AlexNet
Cut precision	.776	.765	.731
Cut recall	.934	.910	.921

Gradual precision	.840	.770	.837
Gradual recall	.904	.622	.386
Cut miss rate	0.066	0.090	0.079
Gradual miss rate	0.096	0.378	0.614
Cut FP ratio	0.289	0.307	0.368
Gradual FP ratio	0.190	0.299	0.195

The three span a wide range, particularly on gradual transitions, where the weakest detector misses nearly two in three. This spread is not a defect of the selection. It reflects the actual state of the tools a researcher might plausibly have used at different points in the last decade, and the point of the exercise is to see what that variation does.

C. Simulation

For each video and each error profile we generate an observed shot list from the true one. At every true boundary, a miss is drawn with the probability appropriate to that boundary's actual transition type, and a missed boundary merges the adjacent shots. Misses are allowed to chain: if the boundary following a merged shot is also missed, the shots merge again. False positives are drawn at the corresponding rate and split a shot at a uniformly random interior point.

One decision here is not forced by the data and should be flagged before the results rather than after them. A false-positive rate is published per true boundary, but merged shots absorb several true boundaries, and how many split opportunities such a shot receives is a modelling choice rather than a measurement. We draw one opportunity per true shot absorbed, which keeps the rate defined per true boundary as published. The alternative, one draw per observed shot, is also defensible, and Section V.B reports both, because the choice turns out to matter for one of the three detectors. We run one thousand iterations per video per profile and compare the mean observed statistic against the true one.

One error source is deliberately absent from the model, and its absence runs in a known direction. The matching criterion described above counts a prediction as correct on a single frame of overlap, so a boundary detected twenty frames from its true position scores as a hit while displacing twenty frames of duration from one shot to its neighbour. Published precision and recall are blind to this. The ClipShots authors evidently recognised as much, since they add a separate intersection-over-union criterion to assess localisation. Our simulation places every detected boundary correctly and models only misses and false positives. Localisation error would add further distortion on top of what we report. The figures below should therefore be read as a floor rather than an estimate.

Three statistics are computed: average shot length, median shot length, and the coefficient of variation. The first two are what the literature reports. The third is included because Redfern's work suggests dispersion carries information that central tendency does not.

For the second question, how large a difference between two videos must be before it survives the error, we take the longest video in the corpus, construct a synthetic comparison video by scaling its shot durations by a fixed ratio, apply error to both, and record how often the comparison recovers the correct ordering.

The analysis code is a single Python file with no dependencies beyond the standard library.

V. RESULTS

A. Shot length is deflated, dispersion inflated

Table 2 gives proportional bias in each statistic under each error profile, averaged across videos. Figures are from one thousand iterations per video per profile under the false-positive model described in Section IV.C.

Table 2. Distribution across the 92 videos of proportional bias in average shot length (%), by detector profile. The 95% range is the 2.5th to 97.5th percentile of per-video bias.

Detector	Mean	Median	SD	95% range
DSM	−15.4	−16.9	3.2	−18.1 to −8.3

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1722842

deepSBD (ResNet-18)	−10.6	−14.9	8.7	−18.0 to +8.9
deepSBD (AlexNet-like)	−2.6	−17.0	29.0	−22.4 to +69.7

All three profiles deflate average shot length, but the summary statistic one chooses matters considerably. For DSM the distribution is tight: a standard deviation of 3.2 percentage points and a 95% range that never crosses zero, so every video in the corpus is deflated. For the AlexNet-based profile the mean of −2.6% is close to meaningless. Its standard deviation is 29 percentage points and its median bias is −17.0%, so the typical video is deflated about as badly as under DSM, and the mean is dragged toward zero by a small number of dissolve-saturated videos that inflate instead. Section V.D identifies exactly which videos those are. Where the two disagree we regard the median as the better description, and readers should treat the mean bias of a weak detector with suspicion.

To see what an error of this size means in practice, consider an effect this field has actually published. Cutting and colleagues reported that shot lengths were scalloped within film acts, with a peak-to-trough magnitude of about 1.1 seconds against a mean shot length of 7.7 seconds, about 14%, and an effect later withdrawn as an artefact of their own normalisation procedure [8]. The deflation we document is of that order. Effects of this size are published in this field, are taken seriously, and have at least once turned out to be method rather than style. The ordering is itself notable: bias is largest for DSM, the strongest of the three detectors, and smallest for the weakest. We return to why in Section VI, since the explanation matters more than the arithmetic.

The mechanism is simple arithmetic, and Figure 1 sets it out. Misses merge shots and push the statistic up. False positives split shots and push it down. Whichever is more frequent wins.

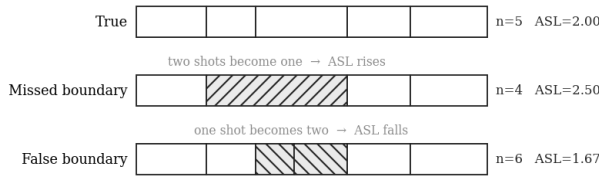

Figure 1. How detection error changes the shot count. A missed boundary merges two shots into one, reducing the count and raising average shot length, while a false boundary splits one shot into two, raising the count and lowering it. Shaded regions mark the affected shots.

On this corpus false positives are more frequent for every detector examined, because precision is the weaker of the two measures across the board. DSM achieves recall above 0.90 on both transition types but precision below 0.85. The result is a systematic underestimate of shot length, which is to say a systematic overestimate of cutting speed.

Median shot length is deflated more than the mean in every case. This is what one would expect once the mechanism is clear: splitting a shot produces two shorter shots and moves mass toward the lower end of the distribution, where the median sits, while the mean is partially defended by the long shots that survive intact. Studies that follow Redfern's recommendation and report the median rather than the mean are, on this evidence, using the more affected statistic rather than the less.

The coefficient of variation rises under every profile, by between 18% and 24%. Detection error adds dispersion regardless of what it does to the centre, since both merging and splitting manufacture durations the film does not contain. Work using dispersion as a stylistic signal should expect it to be inflated.

B. Misses chain, and the estimate is sensitive to how splits are counted

A first-order calculation of the balance between merges and splits, each error rate multiplied by the count of transitions of its type, gives the following per hundred true boundaries:

Table 3. First-order balance of merges and splits per hundred true boundaries.

Detector	Merges per 100	Splits per 100	Net
DSM	7.4	26.3	−19.0
deepSBD (ResNet-18)	16.5	30.5	−14.0
deepSBD (AlexNet-like)	21.8	32.3	−10.5

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1722842

The ordering matches the simulation, but the magnitudes do not, and the discrepancy widens as the miss rate rises. The reason is that misses do not act independently. When a boundary is missed, the next boundary may also be missed, and the observed shot then absorbs three original shots rather than two. If p is the probability of missing a boundary, the number of true shots absorbed into one observed shot follows a geometric distribution with mean $1/(1 - p)$. For DSM, with an effective miss rate of 0.074 across the corpus's mix of transition types, that mean is 1.08, so chains are rare. For the AlexNet-based detector, which misses 61.4% of gradual transitions, the effective rate is 0.218 and the mean run reaches 1.28, with a long tail of much longer runs.

This chaining has a consequence for the simulation that we did not anticipate and that bears on how much confidence the results deserve. A false-positive rate is reported per true boundary, but once shots have merged, it is not obvious how many split opportunities a merged shot should receive. Two defensible answers exist. One draw per observed shot treats the detector as making a bounded number of errors per segment it emits. Draws proportional to the number of true shots absorbed treats the false-positive rate as a property of the underlying footage. We implemented both.

Table 4. Bias in average shot length under the two false-positive counting models.

Detector	One draw per observed shot	Draws per true shot absorbed
DSM	−14.0%	−15.4%
deepSBD (ResNet-18)	−5.4%	−10.6%
deepSBD (AlexNet-like)	+8.1%	−2.6%

For the two stronger detectors the models agree on sign and roughly on magnitude. For the weakest they disagree on sign. This is exactly what the chaining analysis predicts: the two counting models differ only inside merged runs, so they converge when runs are short and separate when runs are long.

We report the second model throughout, on the grounds that a rate measured per true boundary should be applied per true boundary. The deflation result is therefore robust for the stronger of the detectors benchmarked here, and is not robust for a detector missing most of its gradual transitions. Anyone extending this approach should report both counting models rather than settle on one silently: we found that choosing a single model without checking the alternative produced a conclusion the alternative did not support.

C. Small differences are not recoverable

Figure 2 plots how often a comparison between two videos recovers the correct ordering, as a function of the true difference between them, under each error profile.

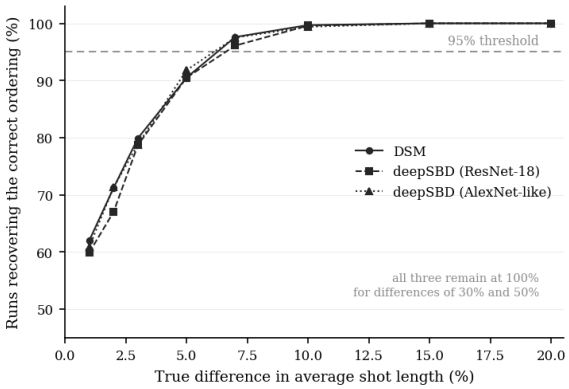

Figure 2. Proportion of simulated comparisons recovering the correct ordering of two videos, as a function of the true difference in average shot length between them. All three error profiles cross the 95% threshold between 5% and 7%.

The three profiles behave almost identically here, which is worth noting in itself: the recoverability of a comparison is far less sensitive to detector quality than the bias in the underlying statistic is. Under the DSM profile, with 1,000 simulated comparisons at each ratio, a true difference of 1% is recovered 61.2% of the time (95% CI 58.2 to 64.2), barely better than a coin. At 3% this rises to 79.5% (77.0 to 82.0) and at 5% to 91.6% (89.9 to 93.3), still short of conventional confidence. At 7% the estimate is 95.5% (94.2 to 96.8), which straddles the 95% mark rather than clearing it, so the crossing point is somewhere in the 5% to 7% band and our data do not locate it more precisely than that. By 10% recovery is 99.7% (99.4 to 100.0).

We suggest this threshold, rather than the bias itself, is the figure worth carrying into practice. Bias can be argued about. A value below which a comparison is uninformative is directly actionable. Under the conditions tested here, two items whose true average shot lengths differ by less than about 5% were not

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1722842

reliably distinguished on that basis from automatically detected boundaries. This figure is a simulation-specific lower bound on recoverable difference rather than a general threshold: the comparison film is constructed by scaling a single real one, so the pair share an identical gradual share and distributional shape, which Section V.D shows is the most favourable case. Section VII sets out why a real comparison would require more. Between 5% and 7% the comparison became defensible but should be presented with the measurement error acknowledged rather than as a clean result.

D. The error is confounded with style

The account in Section II predicts something the results so far have not tested. If the divergence between visual discontinuity and shot change is structured by properties of the footage rather than distributed at random, then bias should not be a single number attaching to a corpus. It should vary from video to video, and vary with whatever property drives the divergence. Gradual transitions are missed between three and six times more often than abrupt ones across the three detectors, so the most obvious candidate is the proportion of a film's transitions that are gradual.

That proportion is not a technical property. It is a stylistic one, a matter of how a film articulates its joins, which varies by period, genre, national tradition and individual practice, and which film scholarship has always treated as meaningful.

We computed each video's gradual share from its annotations and its mean simulated bias under each error profile. The relationship is almost perfectly monotonic: Spearman's rho is +0.98 for DSM, +0.99 for deepSBD/ResNet-18, and +1.00 for the AlexNet-based detector. Figure 3 plots every video, and Table 5 gives mean bias by quartile of gradual share.

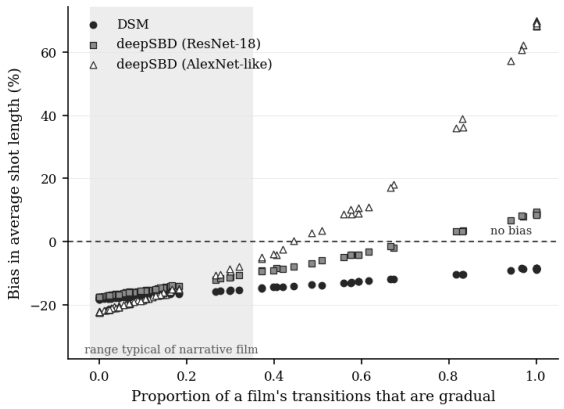

Figure 3. Simulated bias in average shot length for each of 92 videos, plotted against the proportion of that video's transitions that are gradual. Three published error profiles. The shaded band marks the range of gradual shares plausible for narrative feature film. The points to its right are included for completeness but represent material far outside normal practice.

Table 5. Bias in average shot length by quartile of gradual share, with mean gradual share for each quartile in parentheses.

Detector	Q1 (0.01)	Q2 (0.08)	Q3 (0.26)	Q4 (0.81)
DSM	−18.0 %	−17.4 %	−15.8 %	−10.5 %
deepSBD (ResNet-18)	−17.4 %	−16.0 %	−11.9 %	+3.2%
deepSBD (AlexNet-like)	−21.7 %	−19.4 %	−10.3 %	+41.1 %

Two points should be made about this table before anything is concluded from it. The fourth quartile is extreme material: a gradual share of 0.81 means four transitions in five are dissolves or fades, which is not a proportion one meets in narrative feature film, and the +41.1% figure should be read as a demonstration of the mechanism rather than an estimate of anything. The comparison to take seriously is between the first and third quartiles, which span gradual shares of 1% and 26% and do describe a plausible range for real films. Across that range the bias differs by 2.2 percentage points for DSM, 5.5 for deepSBD/ResNet-18 and 11.4 for the AlexNet-based detector.

Even at the low end that is a difference of the same order as the stylistic effects cinemetrics reports as findings. And the direction is consistent: the more a film uses gradual transitions, the less its shot length is deflated, because missed dissolves merge shots and push the measured length back up.

The consequence is not simply that measurement is noisy. It is that the noise is correlated with the variable of interest. Two films that differ in editing style are measured with different bias, in a direction determined by that very difference in style. A hard-cut film and a dissolve-heavy film are not two imperfect measurements of the same quantity with a common error; they are measurements distorted in

opposite directions by the property that distinguishes them. Any comparison between them on shot length therefore includes a contribution from the measurement apparatus that is confounded with the comparison being made.

VI. DISCUSSION

The results bear on a particular kind of claim. Quantitative film studies has a long interest in historical trends in cutting rate: Salt's account of an increasingly restricted stylistic norm [22], Cutting and colleagues' finding of shortening shots across seventy-five years of Hollywood [6], Salt's later argument that the acceleration has stopped [23]. More recently the same framework has been turned on short-form platform video, where average shot lengths are reported to be very short indeed.

Those claims are the ones most exposed here, for two reasons that compound. Fast-cut material has more boundaries per minute, so more opportunities for error. And short-form video is precisely the content on which generic detectors were found wanting badly enough that specialised models were built for it [30]. Our results indicate that a modern detector applied to such material returns an average shot length that is too low, the same direction as the claim being tested. A finding of rapid cutting, obtained with a tool that manufactures rapid cutting, requires more support than the number alone.

We want to be careful about how far this goes. We have not shown that any particular published finding is wrong. The historical trend Cutting and colleagues describe is large, consistent across a substantial corpus [6], and far above the recoverability threshold we identify. Nothing here threatens it. What is threatened is the fine-grained comparison, one director against another, one period against an adjacent one, one platform against another, where the differences at stake are of the same order as the measurement error.

The confounding result in Section V.D sharpens this considerably, and points at the field's most prominent finding. Optical transitions have not been used at a constant rate across film history, and, importantly for what follows, the pattern is not a simple decline. Cutting and colleagues tracked dissolves across 150 Hollywood films released between 1935 and 2005. Fades and wipes became increasingly rare over that period, a finding they report again in later work on narrative discontinuity [9], and dissolves diminished too, but after a lull between 1970 and 1990 they became more numerous again, without approaching studio-era levels [7]. Earlier evidence indicates how gradual-heavy the studio era was: in Carey's sample of 1930s films, dissolves and fades together accounted for 90% of scene transitions, wipes for 9%, and the straight cut for 1% [3], cited in [7].

Set that trajectory against Figure 3. If measured shot length is biased as a function of gradual share, and gradual share across film history is high, then low, then partially recovered, the bias attaching to a decade's films is not constant and not monotonic either. It follows the same U-shape. A study plotting average shot length by decade would therefore carry a decade-varying error term with a minimum somewhere around the 1970s and 1980s.

The concern this raises is not simply that a trend is exaggerated. It is that a non-monotonic bias imposed on a monotonic trend can manufacture curvature, or flatten curvature that is really there. Claims about when cutting accelerated, whether the acceleration was steady or punctuated, and whether it has recently stopped are claims about the shape of a curve, and the shape is what a time-varying bias distorts most readily. We cannot quantify this: our corpus is not feature film, and establishing the effect would require per-decade gradual shares matched against per-decade shot-length measurements from the same films. We can say that the ingredients for the problem are all documented, and that we have found no study combining them.

A related caution applies to genre. Cutting and colleagues' corpus spans action, adventure, comedy, drama and animation, and dissolve usage is not uniform across them. Genre comparisons on shot length therefore inherit the same confound as historical ones.

The confound also operates inside a single film, which Section V.D was not designed to detect. Non-cut transitions are not spread evenly across a running time. Cutting and colleagues report that dissolves, fades and wipes cluster at the openings of films, in the development section and in epilogues, and are rare during the first part of the climax. Montage sequences, which they define as three or more consecutive dissolves, are concentrated in

development and setup and are uncommon in the climax [8]. If bias tracks gradual share, and gradual share varies by act, then the bias is not stationary across a film. It will be heaviest in exactly those passages (openings, montage sequences, transitional development material) where a scholar might be most interested in comparing pacing against the rest of the film. Establishing the size of that within-film variation is beyond what our corpus can support, and we flag it as the most obvious extension of this work.

One further claim of theirs we tested and could not confirm. Cutting and colleagues find that shots flanking a dissolve are long relative to a film's median, which they read as visual preparation for a scene change followed by re-acceleration [7]. Were that true of our corpus, missed gradual transitions would merge already-long shots and inflate average shot length more than our uniform model allows. In ClipShots the median shot flanking a gradual transition is 1.02 times the film median against 1.00 for shots flanking cuts, effectively no difference, although the means diverge more sharply, at 2.03 against 1.49, indicating a heavier upper tail rather than a shift in the typical case. We take this as a reminder that ClipShots is web video and not Hollywood film, and we do not import the finding.

The conceptual argument of Section II can now be restated, since the empirical result gives it force. If detectors merely made mistakes, the appropriate response would be to build better detectors and wait. But the deflation observed here is not a defect that better engineering removes, and Table 2 makes the point more sharply than we expected. The largest bias belongs to DSM, the best of the three detectors by every published F-score. The smallest belongs to the AlexNet-based detector, the worst. In these simulations the detector with the strongest published F-score produced the largest mean bias, and the qualification matters. Measured by the mean, the ordering runs backwards: DSM at −15.4%, the ResNet-18 profile at −10.6%, the AlexNet profile at only −2.6%. Measured by the median it does not run backwards at all. The typical video is deflated by 16.9% under DSM and by 17.0% under AlexNet, which is to say equally.

The discrepancy is the finding. The AlexNet profile misses 61.4% of gradual transitions, and on a handful of dissolve-saturated videos that merging overwhelms its false positives and inflates the statistic sharply, by as much as 70%. Those few videos drag the mean toward zero while leaving the median untouched. The weak detector's apparent advantage is therefore not a cancellation of errors within each video but an artefact of averaging across a corpus whose bias distribution is extremely skewed. Its standard deviation of 29 percentage points against DSM's 3.2 says the same thing.

This carries a practical consequence. Detector quality as conventionally summarised by F-score is therefore not sufficient to predict downstream bias in average shot length, and selecting on F-score alone offers no guarantee of a smaller one. What determines the bias is the ratio of false positives to misses on the particular material, which is not what detector benchmarks are designed to report.

This is why the framing in Section II matters practically and not only philosophically. A field that understands its instrument as an imperfect shot-counter will treat these results as a call for better tools. A field that understands the instrument as a discontinuity-counter used in place of a shot-counter will recognise that some quantity of bias is structural, and will ask instead what claims can survive it. The second response seems to us the correct one, and it is the one that leads to the recoverability threshold in Section V.C rather than to a recommendation that everyone wait for a better model.

A further point concerns reporting. Almost none of the information needed to assess these effects appears in published cinemetric work. Which detector was used is sometimes stated; the threshold or operating point almost never is; validation against any ground truth is rare. Since the size of the bias depends on precisely those details, and not in the direction one would guess, their omission is not a minor gap in reporting practice. A minimal standard would be to name the detector and its settings, and to state whether any validation was performed. That costs an author two sentences.

Finally, an absence we could not resolve. We searched for a publicly available, hand-annotated shot boundary benchmark for feature film and did not locate one, though we should be clear that this was a targeted search rather than a systematic review and we may have missed something. One recent paper reports a 100-minute public test set drawn from ten films [13] but does not appear to distribute it. Of the large movie datasets we examined, none could be

confirmed as carrying human-annotated shot boundaries rather than detector output. If no such benchmark exists, then cinemetrics applies these tools overwhelmingly to the one category of moving image against which they cannot be checked, and the field could close that gap cheaply. If one does exist, we would be glad to be corrected.

VII. LIMITATIONS

The study is a simulation, and its limitations follow from that.

The most consequential is the one described in Section IV.C and Section V.B. How false positives are counted within a merged run is a modelling decision, and for the weakest of the three detectors the two defensible choices give opposite signs. We have reported both rather than selecting one, and we would put no weight on the AlexNet-based figures in isolation. The results for the two stronger detectors are stable across the choice, and it is those we would defend. A reader who prefers the other counting model should read Table 4 rather than Table 2 and will reach the same qualitative conclusion about modern detectors.

The error rates are borrowed rather than measured. They were measured on this corpus, which is the strongest form of borrowing available, but they were measured by the detectors' authors and no independent validation was performed here. Running even one detector against the ground truth ourselves would materially strengthen the study, and we regard that as the obvious next step rather than as an argument against the present one.

The corpus is web and short-form video, not feature film. ClipShots is deliberately difficult material (hand-held, occluded, heterogeneous), and detector performance on it is correspondingly worse than on studio production. The magnitudes reported here should not be transferred to feature film. The underlying arithmetic is general, though its empirical magnitude and direction depend on the detector's error profile and on the characteristics of the material.

We treat the ClipShots annotations as ground truth. They are human annotations and will contain errors of their own, and Song and colleagues have shown that an imperfect gold standard biases the performance figures computed against it [25]. The annotations were produced with double coding for quality assurance, which is better practice than many comparable resources, but no agreement statistic is reported and we cannot quantify the residual uncertainty. It is not reflected in our figures.

The recoverability threshold in Section V.C rests on a narrow base. We take a single video and construct its comparison by scaling every shot duration by a fixed ratio, so the two items being compared share an identical gradual share and an identical distributional shape. That is the most favourable case for recovery, since Section V.D shows bias tracks gradual share: two real films differing in shot length would usually differ in transition style too, and would therefore carry different bias. The 5% figure should be read as a lower bound on what is required, not as a threshold that suffices.

We model misses and false positives but not localisation error, for the reason given in Section IV.C. Since displaced boundaries redistribute duration between adjacent shots while counting as correct detections under the standard matching criterion, this omission means our results understate the total distortion rather than overstating it.

Finally, we do not correct the bias we describe, and the reason is worth stating because it connects to Section VI. Every correction method in this literature (regression calibration, multiple imputation, the maximum likelihood adjustment of TeBlunthuis and colleagues) requires gold standard validation data: a subset of cases where both the classifier's output and the true human-coded value are known. The corrections are only as good as that validation set.

For cinemetrics on feature film, no such validation set exists publicly. That is the same absence noted in Section VI, arriving from the other direction. The field cannot correct for detection error in the material it principally studies, because it has never assembled the ground truth that a correction would have to be calibrated against. Establishing the size and direction of the bias is therefore not merely a first step we happened to take. It is as far as anyone can currently go. Building the benchmark is the precondition for everything else.

VIII. CONCLUSION

Automatic shot detection made quantitative film studies cheap, and the field has taken up the tools with more enthusiasm than scrutiny. The error rates of those tools are not secret. They are published, sometimes in the same documents that distribute the

data, and they are large enough to move the statistics reported by margins comparable to the stylistic differences being claimed.

Under simulations calibrated to published ClipShots performance, the two stronger detector profiles produce a downward bias of 10% to 15% in average shot length, and more in median shot length. The material is made to look as though it cuts faster than it does. That this runs in the same direction as the field's most prominent recent claims about accelerating cutting rates is not evidence that those claims are wrong, but it does mean they cannot be supported by detector output alone.

The sharpest form of the problem is that this error is not independent of what the field wants to know. Bias tracks a film's use of gradual transitions almost perfectly, and the use of gradual transitions is a stylistic choice. The instrument distorts films in proportion to a property that film scholarship has always regarded as significant, which means the distortion cannot be treated as background noise to be averaged away across a large corpus. It travels with the signal.

Underneath the arithmetic is a substitution the field has made without much discussion. The shot was a concept applied by a person. It is now a discontinuity registered by an instrument. The substitution was reasonable, it was necessary if the work was to be done at scale, and most of the time the two coincide. But they diverge in ways that are systematic rather than random, and a statistic computed across thousands of boundaries accumulates systematic divergence rather than averaging it away. What cinemetrics now measures is not quite what it says it measures, and the gap has a direction.

The modest practical recommendation is that authors state which detector they used and at what settings, and treat differences in average shot length below about 5% as uninformative unless they have validated their boundaries. The less modest one is that the field acquire a hand-annotated feature film benchmark, so that the question of how far these tools can be trusted on the material they are mostly used for can be answered rather than assumed. Neither recommendation requires anyone to stop counting shots. They require only that the field say what it is counting.

Data and code availability

The ClipShots annotations are distributed by their authors at https://github.com/Tangshitao/ClipShots under an MIT licence covering the repository contents. We used only the annotation files, and no video was downloaded or redistributed. The underlying videos are third-party material sourced from YouTube and Weibo, and no licence covering them is stated by the dataset authors.

Analysis code is a single Python file (`propagate.py`) using only the Python standard library, so it runs without installing anything. Figure generation (`figures.py`) additionally requires matplotlib. Both are available from the corresponding author on request.

REFERENCES

[1] Bachl, M., & Scharkow, M. (2017). Correcting measurement error in content analysis. Communication Methods and Measures. https://doi.org/10.1080/19312458.2017.1305103.

[2] Bakels, J.-H., Grotkopp, M., Scherer, T., & Stratil, J. (2020). Matching computational analysis and human experience: performative arts and the digital humanities. Digital Humanities Quarterly, 14(4). https://doi.org/10.17169/refubium-30938.

[3] Chávez Heras, D. (2024). Creanalytics: automating the supercut as a form of critical technical practice. Convergence: The International Journal of Research into New Media Technologies, 30(1), 264–284. https://doi.org/10.1177/13548565231174592.

[4] Chen, B. (2025). Cinemetrics in the digital humanities era: theoretical evolution, methodological innovation, and research frontiers. The Development of Humanities and Social Sciences. https://doi.org/10.71204/dxaqff27.

[5] Cutting, J.E., Brunick, K.L., & DeLong, J.E. (2011). How act structure sculpts shot lengths and shot transitions in Hollywood film. Projections, 5(1), 1–16. https://doi.org/10.3167/proj.2011.050102.

[6] Cutting, J.E., Brunick, K.L., DeLong, J.E., Iricinschi, C., & Candan, A. (2011). Quicker, faster, darker: changes in Hollywood film over 75 years. i-Perception, 2(6), 569–576. https://doi.org/10.1068/i0441aap.

[7] Cutting, J.E., Brunick, K.L., & DeLong, J.E. (2011). The changing poetics of the dissolve in Hollywood film. Empirical Studies of the Arts, 29(2), 149–169. https://doi.org/10.2190/em.29.2.b.

[8] Cutting, J.E., Brunick, K.L., & DeLong, J.E. (2012). On shot lengths and film acts: a revised view. Projections, 6(1). https://doi.org/10.3167/proj.2012.060106.

[9] Cutting, J.E. (2014). Event segmentation and seven types of narrative discontinuity in popular movies. Acta Psychologica, 149, 69–77. https://doi.org/10.1016/j.actpsy.2014.03.003.

[10] Hanjalic, A. (2002). Shot-boundary detection: unraveled and resolved? IEEE Transactions on Circuits and Systems for Video Technology. https://doi.org/10.1109/76.988656.

[11] Kar, T., & Kanungo, P. (2024). Video shot-boundary detection: issues, challenges and solutions. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-10742-1.

[12] Li, C., Lu, J., Pei, Y., Shen, Y., Hu, Y., Fan, Y., Tian, Y., Linghu, X., Wang, K., Shi, Z., & Sun, J. (2024). PyCinemetrics: computational film studies tool based on deep learning and PySide2. SoftwareX, 26, 101686. https://doi.org/10.1016/j.softx.2024.101686.

[13] Li, C., Wang, K., Pei, Y., Fan, Y., & Zhang, F. (2026). FilmShots: a fast film shot boundary detection based on dilated Conv3D and attention. IEEE Access, 14, 49497–49510. https://doi.org/10.1109/access.2026.3677190.

[14] Lienhart, R. (1998). Comparison of automatic shot boundary detection algorithms. https://doi.org/10.1117/12.333848.

[15] Pustu-Iren, K., Sittel, J., Mauer, R., Bulgakowa, O., & Ewerth, R. (2020). Automated visual content analysis for film studies: current status and challenges. Digital Humanities Quarterly, 14(4). https://doi.org/10.63744/8rc7hctnsxrx.

[16] Qiao, M., & Huang, K. (2025). Correcting measurement error in regression models with variables constructed from aggregated output of data mining models. MIS Quarterly, 49(1), 29–60. https://doi.org/10.25300/misq/2024/18026.

[17] Redfern, N. (2014). Comparing the shot length distributions of motion pictures using dominance statistics. Empirical Studies of the Arts, 32(2). https://doi.org/10.2190/em.32.2.g.

[18] Redfern, N. (2015). The log-normal distribution is not an appropriate parametric model for shot length distributions of Hollywood films. Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqs066.

[19] Redfern, N. (2022). Distributional thinking about film style. Exchanges: The Interdisciplinary Research Journal, 10(1). https://doi.org/10.31273/eirj.v10i1.853.

[20] Redfern, N. (2022). Analysing motion picture cutting rates. Wide Screen, 9(1). ISSN 1757-3920. http://widescreenjournal.org.

[21] Salt, B. (1974). Statistical style analysis of motion pictures. Film Quarterly, 28(1), 13–22. https://doi.org/10.2307/1211438.

[22] Salt, B. (2004). The shape of 1999. New Review of Film and Television Studies. https://doi.org/10.1080/17400300410001679106.

[23] Salt, B. (2021). The end of the Great Speed-Up - and after. Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqab049.

[24] Smeaton, A.F., Over, P., & Doherty, A.R. (2010). Video shot boundary detection: seven years of TRECVid activity. Computer Vision and Image Understanding. https://doi.org/10.1016/j.cviu.2009.03.011.

[25] Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H.G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. Political Communication, 37(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752.

[26] Tang, S., Feng, L., Kuang, Z., Chen, Y., & Zhang, W. (2018). Fast video shot transition localization with deep structured models. arXiv:1808.04234.

[27] TeBlunthuis, N., Hase, V., & Chan, C.-H. (2023). Misclassification in automated content analysis causes bias in regression. Can we fix it? Yes we can! Communication Methods and Measures, 18(3). https://doi.org/10.1080/19312458.2023.2293713 Preprint: arXiv:2307.06483.

[28] Wang, S., McCormick, T.H., & Leek, J.T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. PNAS. https://doi.org/10.1073/pnas.2001238117.

[29] Wu, X. (2026). Kurosawa shot-level dataset: shot boundaries, color, and face-derived shot-scale features for 30 films. Journal of Open Humanities Data, 12. https://doi.org/10.5334/johd.507.

[30] Zhu, W., Huang, Y., Xie, X., Liu, W., Deng, J., Zhang, D., Wang, Z., & Liu, J. (2023). AutoShot: a short video dataset and state-of-the-art shot boundary detection. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2238–2247. https://doi.org/10.1109/cvprw59228.2023.00218.