VIDEO NEWS SHOT LABELING REFINEMENT VIA SHOT RHYTHM MODELS

John R. Kender

Department of Computer Science
Columbia University
New York, NY 10027
jrk@cs.columbia.edu

Milind R. Naphade

IBM T J Watson Research Center
Business Informatics Department
Hawthorne, NY 10532
naphade@us.ibm.com

ABSTRACT

We present a three-step post-processing method for increasing the precision of video shot labels in the domain of television news. First, we demonstrate that news shot sequences can be characterized by rhythms of alternation (due to dialogue), repetition (due to persistent background settings), or both. Thus a temporal model is necessarily third-order Markov. Second, we demonstrate that the output of feature detectors derived from machine learning methods (in particular, from SVMs) can be converted into probabilities in a more effective way than two suggested existing methods. This is particularly true when detectors are errorful due to sparse training sets, as is common in this domain. Third, we demonstrate that a straightforward application of the Viterbi algorithm on a third-order FSM, constructed from observed transition probabilities and converted feature detector outputs, can refine feature label precision at little cost. We show that on a test corpus of TRECVID 2005 news videos annotated with 39 LSCOM-lite features, the mean increase in the measure of Average Precision (AP) was 4%, with some of the rarer and more difficult features having relative increases in AP of as much as 67%.

1. INTRODUCTION

News stories typically are reported as separate video episodes reported over time and over different channels, with each episode comprised of a sequence of related video shots. One first step to the effective indexing and retrieval of all related video episodes, across all times and all channels, is the annotation of their individual shots using concept tags derived from a formal ontology of visual features. A number of efforts have been made to derive and evaluate such ontologies, with perhaps the most advanced being the the Large Scale Concept Ontology for Multimedia Understanding (LSCOM); see [1] for an overview.

However, nearly all current work on video shot labeling assumes the temporal independence of shots. Shots are annotated in isolation, or, equivalently, as if they had been temporally scrambled. Nevertheless, there are good reasons to suspect that temporal order is significant, and that it may give exploitable cues for more accurate labeling. For example, there are known limits on human visual info processing that tend to bias the selection and editing of shots so that a particular visual-temporal "texture" is preserved throughout a video episode. Further, the economics of video news production tend to limit editorial freedom; the reuse of nearly-identical shots and shot sequences across widely separated episodes have been noted and tracked by a number of researchers.

The work reported here is inspired by the observation that in related video formats such as drama and comedy there has been ample recognition of shot rhythms; for example, see [2]. Researchers have noted the prevalence of video dialogue sequences, in which there is a nearly strict alternating rhythm between two different shots types, but where within each type the shots are nearly identical in foreground and background. Likewise, action sequences have been noted as being unusually diverse and almost random in their feature sequences, deliberately conveying drama by their lack of a steady rhythm. Both these sequences vary greatly from what appears to be the implicit default shot sequence rhythm of a steady repetition of most features, particularly in backgrounds. These three rhythms— alternation, repetition, and randomness—intuitively are also observable in video news episodes. They present an opportunity for improving the automatic labeling of shots, by selectively reinforcing detector response based on their rhythmic reoccurrences.

The analysis and results that follow are based on a 62K-shot subset of the 80 hours of news video annotated as part of the TRECVID Video Annotation Forum of 2005, by over 100 participants via the web-based Efficient Video Annotation (EVA) System [3], using a 39-feature ontology called

This work was completed while the first author was supported at IBM Research. This material is based upon work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

LSCOM-lite. This ontology was derived in part by analyzing of the strengths and weaknesses of a predecessor 133-feature ontology used in a similar TRECVID effort in 2003. The LSCOM-lite ontology appears to be better tuned for the task; for example, the frequencies of annotation of the 17 most common features (Person, Face, Outdoor, Crowd, Studio, Sky, Entertainment, Building, Walking/Running, Vegetation, Car, Government-Leader, Urban, Road, Meeting, Military, Computer/TV-Screen) appear to closely follow Zipf's law, just as do more established vocabularies such as the full English language.

Nevertheless, the automated labeling of such concepts remains inexact, and even the measures of quality of labeling are controversial. One very conservative measure is Average Precision (AP), defined as the average, over the size of the ground truth, of the instantaneous precisions of a sequence of experiments. Each experiment retrieves new candidate shots one by one until a new correctly labeled shot is found. Instantaneous precision is then defined as the number of correctly labeled shots (which increases by exactly one at each step) divided by the total retrievals in all experiments so far (which includes all the errors of this and all prior experiments). Early errors of retrieval therefore continue to penalize subsequent experiments severely.

Some features, such as the three most common ones, can be detected in isolated key frames with much greater than 90% AP. However, AP quickly drops as features become less common, in part because less training data is available. For example, the AP for Building is typically less than 50%, and most of the rarer concepts, such Police-Security or Prisoner, typically have an AP in the low single digits of percent. This work presents a low-cost post-processing method for increasing AP by exploiting feature-dependent shot rhythms. On average, it increases absolute AP by 4%, although particular features that follow more pronounced rhythms see increases of relative AP as much as 67%.

2. SHOT TEMPORAL RELATIONSHIPS

We examined the 62K annotated shots and noted that probabilities of occurrence of a feature were neither temporally independent nor (first-order) Markovian. This was true even though this analysis was a bit sloppy, in that all shots were concatenated together into a single sequence without regard to episode boundaries; this introduced (a relatively small number of) spurious transitions. We found that features that indicate the presence of a particular physical setting (for example, Sports or Entertainment) tend toward repetitions; their probability of reoccurrence in a second shot following a first one with this feature approaches .9. In contrast, features that indicate the presence of an object reoccur only with a probability of about .3. Over all features, the probability of immediate second recurrence is from 2 to 700 times more likely that the probability of first occurrence. Additionally, for features that tend to appear in dialogue sequences (for example, Person), the reoccurrence probabilities are greater still at the *third* shot, and are there as much as 250 times higher than first occurrence probabilities.

We examined all conditional probabilities reflecting dependencies on sequences of up to six shots, but found through a principal component decomposition of these transition matrices that only repetition (first-order Markov) and alternation (second order Markov) were significantly represented. We also noted that alternation was more clearly evident when a history of three prior states were kept (third order Markov), as this more clearly distinguished between alternation and sporadic insertion/deletions. That is, sequences of characteristic feature presence of 1010 and 0101 are clearly alternating, but the sequences of 010 and 101 may simply indicate a noisy 1 or 0, respectively. A third order Markov chain for feature presence has 8 states and 16 transitions, but because of various conservation rules, one can show that it has only 7 free parameters, which simplifies data-gathering. Additional investigation also found that alternating features did not prefer to alternate with any particular class of feature; those statistics were independent.

Using the first two eigenvectors of their transition probability vectors, we show in Fig. 1 the 39 features as they are located in this two dimensional space of repetition versus alternation. It is apparent that this space is a continuum and that most features are mixed.

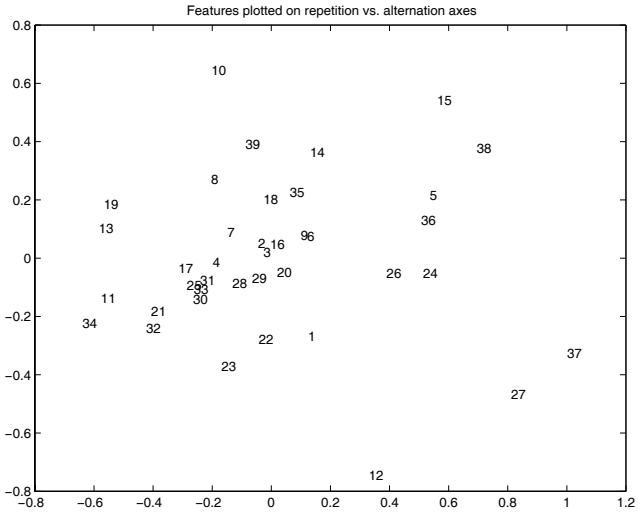

Fig. 1. A display of the likelihood of feature repetition (horizontal) vs. feature alternation (vertical), showing these properties of the 39 LSCOM-lite features. Features describing backgrounds (37=Weather, 27=Court) tend toward repetition; features describing human interactions (15=Person, 14=Face, 39=Meeting) tend toward alternation.

3. FROM SVM SCORES TO PROBABILITIES

The training set of 62K shots easily provides, for each feature, the transition probabilities for its recognizing FSM. However, before we could run the usual Viterbi algorithm on the shot sequences to refine existing feature scores, we needed to convert raw feature values into raw feature probabilities. We explored two existing popular methods for doing so.

Our feature detectors are Support Vector Machines, which by definition return for each shot a score, s, which measures the (hyperspace) distance of the input shot vector from the hyperplane decision margins. SVMs are calibrated so that a value of $s = +1$ occurs at the positive margin, and $s = -1$ occurs at the negative margin. Monotonically increasing scores indicate monotonically increasing certainties of classification.

The method of Platt is based on an empirical observation that suggests fitting these scores to a two-parameter family of curves, $1/(1 + \exp(As + B))$ that intuitively capture a common probability measure ("log odds") [4]. However, as our data shows in Fig. 2, and as the paper of Zadrozny [5] more powerfully demonstrates, this family of curves often poorly captures the underlying probabilities, even when SVM scores are normalized to fall within a pre-specified range.

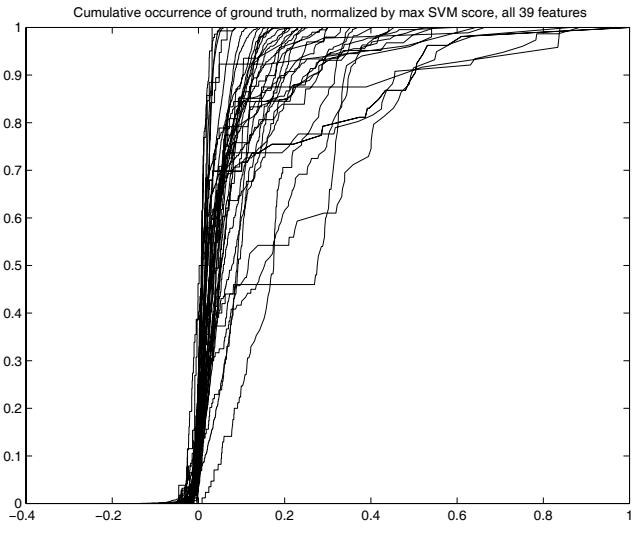

Fig. 2. For each feature, SVM score normalized by maximum score (horizontal), versus cumulative occurrences of ground truth normalized by total occurrence (vertical). No simple probability model fits, especially for rarer features.

Unfortunately, the alternative method of isotonic regression suggested in [5] doesn't appear to help either, particularly for rarer features. With limited training data, SVM responses for these features are errorful, with many positive scores corresponding to negative ground truth. Therefore, the attempt to estimate feature probability by effectively finding local ground truth density (subject to a side condition of monotonic growth) often ends up severely underestimating true probabilities, particularly near the positive margin, as Fig. 3 shows. These diminished probabilities make FSM refinement overly cautious: few features are detected and reinforced, and AP is improved on average by only 1%, with no feature improving more than 5%.

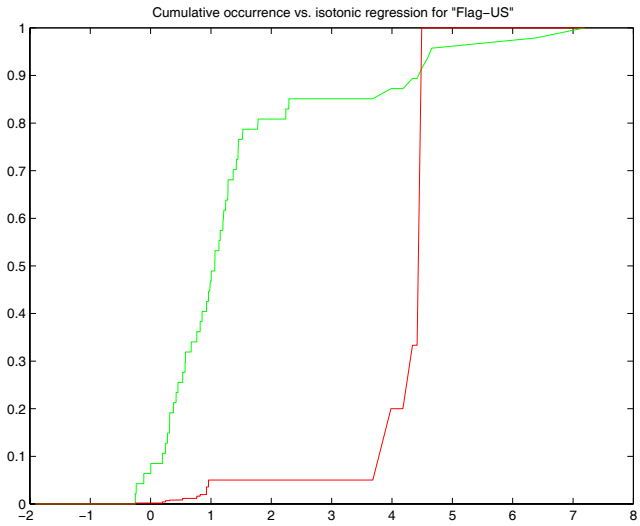

Fig. 3. Isotonic regression (lower curve) severely underestimates feature probability for sparse features, especially near positive margin. But normalized cumulative occurrence compensates for SVM errors at all positive scores.

What we have found instead, to empirically compensate for the poor SVM performance on sparse features, is the use of a normalized cumulative occurrence measure of the ground truth set G; see again Fig. 3. We define the value of this measure at s as the fraction of total ground truth positives p that have a SVM score less than or equal to s. That is,

$$cdf(s) = |p : p \subset G \wedge score(p) <= s|/|p \subset G|$$

For more popular features, this measure tracks the isotonic regression curve well; for sparser ones it accumulates more history than isotonic regression allows and therefore better interpolates over gaps in the positive SVM response. And, since in the neighborhood of the separating hyperplane the SVM response curve appears to reflect an exponential probability density function [4], the normalized cumulative measure, its integral, will be exponential as well.

4. PERMUTING THE SCORE RANKINGS

Once we have an FSM and feature probabilities, it is straightforward to run the Viterbi algorithm to refine which shots had a given feature. We trained on 62K shots of data and tested on 6.5K shots. We found that performance improved somewhat by using a separate normalized cumulative measure for feature absence, also; these two cumulative measures do not quite add to 1 at every score s.

To measure the effectiveness of this procedure, we used AP again, feeding it shots in a permuted order. We first fed the AP algorithm only those shots that had been detected by Viterbi, and in order of their declining raw SVM scores. After these shots were exhausted and, if it was still necessary (we used the customary cut-off of 1000 shots), the remaining shots were fed to the AP algorithm, again in declining score order. The AP computation of a feature having a strong rhythm is shown in Fig. 4.

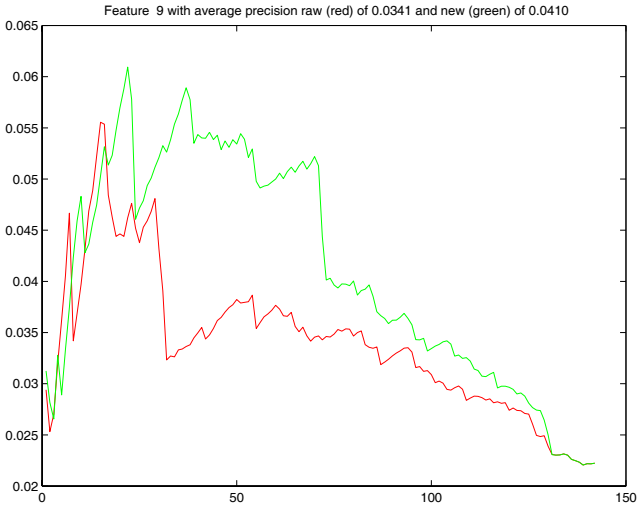

Fig. 4. A graph of the improvement in the Average Precision of the highly alternating feature Corporate-Leader (upper curve): horizontal axis is experiment number, vertical axis is instantaneous precision at that experiment.

Overall, the average absolute improvement in the highly conservative AP measure was 4%, above average baseline absolute AP measures of about 30%. (That is, performance increased from about 30% to about 34%.) In general, the more pronounced was a feature's shot rhythm—the closer it fell to either the upper margin or right margin of Figure 1—the more significant was the improvement in AP. We also noted that a first order Markov model was not as successful; it showed a mean increase in AP of only 2.6%.

The one and only striking failure in performance, showing a loss in absolute AP of 17%, was the Computer/TV-Screen feature, a rare feature which nevertheless showed high repetition in the training set, with probability of recurrence of .85. The test set, which may be from a different annotator, displayed radically different transition statistics; its probability of recurrence was only .37.

5. CONCLUSIONS AND FUTURE WORK

We presented a three-step post-processing method for improving the precision of feature detection in shots, based on derived shot rhythms. Since the method only visits the derived feature scores and not the images themselves, its cost is very low, and linear in the number of shots. It necessarily uses a third order Markov model. It shows good results even though the training was somewhat sloppy and the ground truth annotations appear to have a number of serious errors.

In future work, we anticipate: improving the training by accommodating episode boundaries; improving the testing by using cross-validation; exploring a rigorous theoretic justification for the cumulative occurrence measure; determining whether the statistics of a given feature can accurately choose its refining FSM's Markovian order—including zeroth order; and applying this shot label refinement method to concrete indexing and retrieval applications.

6. REFERENCES

[1] Alexander G. Hauptmann, "Towards a large scale concept ontology for broadcast video," in *Proceedings of International Conference on Image and Video Retrieval (CIVR'04)*, July 2004, pp. 674–675.

[2] Brett Adams, Chitra Dorai, and Svetha Venkatesh, "Study of shot length and motion as contributing factors to movie tempo," in *Proceedings of the ACM International Conference on Multimedia (ACMMM'00)*, October 2000, pp. 353–355.

[3] Timo Volkmer, John R. Smith, and Apostol Natsev, "A web-based system for collaborative annotation of large image and video collections: An evaluation and user study," in *Proceedings of the ACM International Conference on Multimedia (ACMMM'05)*, October 2005, pp. 892–901.

[4] John C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, Peter J. Bartlett, Bernhard Scholkopf, Dale Schuurmans, and Alex J. Smola, Eds., chapter 5, pp. 61–74. MIT Press, October 2000.

[5] Bianca Zadrozny and Charles Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD'02)*, July 2002, pp. 694–699.