GENERATIVE, DISCRIMINATIVE, AND ENSEMBLE LEARNING ON MULTI-MODAL PERCEPTUAL FUSION TOWARD NEWS VIDEO STORY SEGMENTATION

Winston H.-M. Hsu and Shih-Fu Chang

Dept. of Electrical Engineering, Columbia University, New York
{winston, sfchang}@ee.columbia.edu

ABSTRACT

News video story segmentation is a critical task for automatic video indexing and summarization. Our prior work has demonstrated promising performance by using a generative model, called Maximum Entropy (ME), which models the posterior probability given the multi-modal perceptual features near the candidate points. In this paper, we investigate alternative statistical approaches based on discriminative models, i.e. Support Vector Machine (SVM), and Ensemble Learning, i.e. Boosting. In addition, we develop a novel approach, called BoostME, which uses the ME classifiers and the associated confidence scores in each boosting iteration. We evaluated these different methods using the TRECVID 2003 broadcast news data set. We found that SVM-based and ME-based techniques both outperformed the pure Boosting techniques, with the SVM-based solutions achieving even slightly higher accuracy. Moreover, we summarize extensive analysis results of error sources over distinctive news story types to identify future research opportunities.

1. INTRODUCTION

News video story segmentation is a fundamental step in news video indexing and understanding. A news story is a basic unit for browsing, summarization, and understanding. Automatic segmentation of continuous video programs into constituent story units is challenging due to the diversity of story types and the complex composition of attributes in various types of stories. Review of existing approaches and definition of news stories can be found in [1].

In our prior work [1], we reported a system that used Maximum Entropy to fuse 195 features from multiple modalities (audio, visual, and/or text), discover salient features, and demonstrate promising performance over the TRECVID 2003 data set (for CNN channel, F1[1]=0.69 using audio-visual features only, F1=0.73 using audio-visual-text features). It models the posterior probability of a candidate point in time to be a true story boundary given the observations of multi-modal multi-level features near the point.

In this paper, we compare fusion capabilities of generative, discriminative, and ensemble learning models on multi-modal feature sets. Our goal is to explore the potential of emerging learning techniques in feature fusion and thereby news story segmentation. In the generative model, we adopt the prior ME approach which has shown encouraging fusion capability in natural language processing [2] and multi-modal fusion [1]. For the discriminative model, we apply SVMs which have demonstrated discriminative power in diverse problems [3]. For the ensemble learning, we applied variations of boosting approaches, which have been used to aggregate weak classifiers [4, 5, 6] efficiently and effectively. In addition, we also develop a novel approach, called BoostME, which combines boosting and ME; the ME classifier is used in each boosting iteration with associated confidence scores while the parameters and features of ME are estimated using the re-weighted training data after each boosting iteration.

Another objective of our study is to obtain a better understanding of the problem by categorizing different production styles of video stories, analyzing the sources of classification errors over different categories, and assessing the strengths and weaknesses of each statistical model against different types of stories. Our findings indicate that SVM-based and ME-based approaches outperform the pure boosting approach, while the SVM methods achieving a slight gain ($F1 = 0.71$ using A+V features only) over the ME methods. The ME-Boosting combined methods achieve about the same performance as the ME methods.

Data set and the extracted feature pool are briefly reviewed in Section 2. The learning approaches are described in Section 3. The experiments and discussions are included in Section 4 and followed by the conclusion and future work in Section 5.

2. DATA SET AND FEATURE POOL

2.1. Data set and candidate points

Like our prior work in [1], we use 218 half-hour ABC World News Tonight and CNN Headline News broadcasts recorded by the Linguistic Data Consortium (LDC) from late January 1998 through June 1998. The video is in MPEG-1 format and is packaged with associated files including automatic speech recognition (ASR) transcripts and annotated story boundaries. The data are prepared for TRECVID 2003[2] with the goal of promoting progress in content-based video retrieval via open metric-based evaluation.

The story boundaries defined by LDC include those of normal news stories as well as boundaries of sports and weather. Figure 1 illustrates common types of stories that can be found in broadcast news videos such as CNN. The proportion of different types in the whole collection is listed in Table 1 (row 3: percentage). Note that there are a broad range of story types with significant percentage of data.

In this experiment, we take the union of shot boundaries and audio pauses as candidate points but remove duplications within a 2.5-second fuzzy window. Our prior study showed these two sets of points account for most of the story boundaries in news [1].

2.2. Raw multi-modal features and feature wrapper

We adopt the raw audio-visual features developed in [1]. They cover a wide range of features at different levels from audio, speech, and visual modalities. They include anchor face, commercial, pitch

[1] $F1 = \frac{2 \cdot P \cdot R}{P+R}$, where P and R are precision and recall rates.

[2] TRECVID 2003: http://www-nlpir.nist.gov/projects/tv2003/tv2003.html

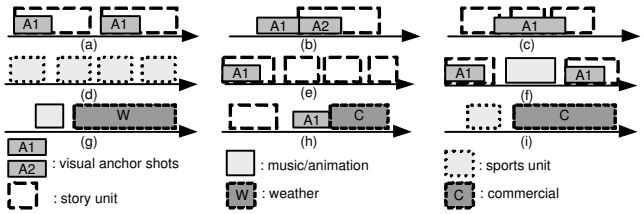

Fig. 1. Common CNN story types seen in the TRECVID 2003 data set. $A1$ and $A2$ represent segments showing visual anchor persons; (a) two stories both starting with the same visual anchor; (b) the second story starts with a different visual anchor; (c) multiple stories reported in a single visual anchor shot; (d) a sports section constitutes series of briefings; (e) series of stories that do not start with anchor shots; (f) two stories that are separated by long music or animation representing the station id; (g) weather report; (h) a non-story section consists of a anchor lead-in followed by a long commercial; (i) a commercial comes right after a sports section.

jump, significant pause, speech segment and rapidity, music/speech discrimination, motion, etc. Details are in [1].

The audio-visual features described in the above section are usually diverse and asynchronous (e.g., features computed at points from surrounding time windows from different modalities). We have developed a feature wrapper to convert different types of features into a consistent form of binary features $\{g_j\}$ [1]. The feature wrapper function also contains parameters of time interval for measuring the change over time, thresholds for quantizing continuous feature values to binary ones, and the size of the observation window surrounding the candidate points. From the raw feature pool described above, we use the feature wrapper to generate a 195-dimension binary feature vector at each candidate point. Both the original raw feature vectors and the wrapped binary feature vectors are evaluated in this study.

3. LEARNING APPROACHES

We include three families of statistical models for comparison in this study - ME, SVM, and Boosting. As mentioned earlier, each of them has been shown with success in solving practical problems. However, they represent interesting and distinctive categories of statistical approaches - generative vs. discriminative vs. ensemble learning. One of our goals is to gain comparative understanding of strengths of each family through the test case of news video segmentation.

3.1. Maximum Entropy approach (ME-BIN-35)

The ME model [1, 2] constructs an exponential log-linear function that fuses multiple binary features to approximate the posterior probability of an event (i.e., story boundary) given the audio, visual, or text data surrounding the point under examination, as shown in Equation 1. The construction process includes two main steps - parameter estimation and feature induction.

The estimated model, a posterior probability $q_\lambda(b|x)$, is represented as

$$q_\lambda(b|x) = \frac{1}{Z_\lambda(x)} \exp\{\sum_i \lambda_i f_i(x,b)\}, \qquad (1)$$

where $b \in \{0,1\}$ is a random variable corresponding to the presence or absence of a story boundary in the context x; $\{\lambda_i\}$ is the estimated real-valued parameter set; $\sum_i \lambda_i f_i(x,b)$ is a linear combination of binary features; $Z_\lambda(x)$ is a normalization factor.

Meanwhile, x is the video and audio data surrounding a candidate point of story boundaries. From x we compute a set of binary features, $f_i(x,b) = 1_{\{g_i(x)=b\}} \in \{0,1\}$. $1_{\{\cdot\}}$ is an indication function; g_i is a predictor of story boundary using the i'th binary feature, generated from the feature wrapper. f_i equals 1 if the prediction of predictor g_i equals b, and is 0 otherwise.

Parameter estimation: the parameters $\{\lambda_i\}$ are estimated by minimizing the Kullback-Leibler divergence measure between the estimated model and the empirical distribution in the training set. We use an iterative process to update $\{\lambda_i\}$ till divergence is minimized. Thanks to the convex property of the distribution, the iterative process is able to find the global optimal parameters.

Feature induction: from the candidate pool, a greedy induction process is used to select the feature that has the largest improvement in terms of gains or divergence reduction. The selected feature is then removed from the candidate pool. The induction process iterates with the new candidate set till the stopping criterion is reached (e.g., upper bound of the number of features or lower bound of the gain). In our experiment, we select 35 binary features.

3.2. Boosting approaches

Boosting techniques have been successfully used to improve classification performance by fusing multiple weak classifiers. In [4], classifiers with binary output are fused to form a linear combination of individual prediction results. In [6], boosting is used to select features from a very large pool for rapid image retrieval. In [5], the real-valued weak classifiers are selected and combined in each iteration to minimize an exponential loss function. Based on these, we have developed three boosting-based methods.

Let $S = \{(x_1, y_1), ..., (x_N, y_N)\}$ be a sequence of training examples x_i with corresponding labels $y_i \in \{-1, +1\}$. At iteration t, $h_t : x \rightarrow [-1, +1]$ is the weak classifier estimated in this step and $h_t(x_i)$ reports the prediction when sample x_i is used as input. Besides, $D_t(i)$ is the weight of training sample x_i at iteration t and is initialized to be $1/N$.

3.2.1. Boosting with binary features (BST-BIN-35)

The approach is very similar to that in [6] except that $h_t = g_{j^*}$, where g_{j^*} is the binary predictor (feature) incurring the least weighted prediction error in iteration t. Note that the search space of j is the entire pool of binary features $\{g_j\}$; e.g., 195 binary features in our experiment.

In comparison with ME induction process, the boosting process selects 35 most salient features after the 35 iterations, each feature is used as a simple binary predictor.

3.2.2. BoostME: Boosting with confidence-rated ME models (BST-ME-Z)

The feature induction method built on the ME approach (Section 3.1) is effective in selecting the most salient feature set from a large pool. The optimization is subject to the relative importance assigned to each training sample. When the training sample weights change, the optimal feature set change accordingly also. Such characteristics might match the strength of the boosting approach - erroneous samples will be emphasized through re-weighting in different iterations. In each iteration of boosting, classifiers with different feature sets can be built to target the erroneous samples

Table 1. Percentages of missed story boundaries (rows 3-7) by different approaches over the TRECVID CNN data set. The error percentages are broken into different types of video stories (defined in Figure 1). The detection precision is fixed at 0.71. Rows 1 and 2 are the number of story boundaries and their percentages. The notations for each method are defined in corresponding subsections in Section 3.

#	Exps./Types	a	b	c	d	e	f	g	h	i	all
1	Story Bdry. #	244	48	67	114	162	16	22	58	28	759
2	Percentage (%)	32.0	6.3	8.8	15.0	21.3	2.1	2.9	7.6	3.7	100
3	ME-BIN-35	4.9	68.8	20.9	59.6	61.7	93.8	27.3	17.2	39.3	35.4
4	BST-BIN-35	6.2	89.6	62.7	83.3	84.0	93.8	22.7	15.5	75.0	50.2
5	SVM-RAW-33	4.9	70.8	19.4	67.5	58.0	93.8	31.8	19.0	21.4	35.4
6	SVM-BIN-35	2.9	66.7	16.4	59.6	56.8	93.8	31.8	12.1	28.6	32.5
7	SVM-BIN-195	2.9	45.8	14.9	54.4	51.2	93.8	36.4	10.3	10.7	28.4

remaining from earlier iterations. Based on this observation, we develop a combined technique, called *BoostME*, which uses ME plus feature induction in each boosting iteration.

For each iteration $t = 1, ..., T$, the BoostME algorithm iterates as the following:

- Induce and estimate an ME model q_λ give the current sample weight $D_t(\cdot)$.
- Choose $h_t(x) = h'(x, d_t^*)$, where $h'(x, d) = \chi(q_\lambda(x), d)$ is the shifted and normalized classification score from the ME model q_λ using threshold d. d_t^* is the best threshold to minimize the expected exponential loss Z_e [5] parameterized with decision threshold d,

$$Z_e(d) = \sum_{i=1}^{N} D_t(i) \exp\{-\alpha'(d) y_i h'(x_i, d)\}; \quad (2)$$

$$\chi(a, d) = 1_{\{a \geq d\}} \cdot \frac{a-d}{1-d} + 1_{\{a<d\}} \cdot \frac{a-d}{d}, a, d \in (0, 1),$$

$$\alpha'(d) = \frac{1}{2} \ln \left(\frac{1 + \sum_i D_t(i) y_i h'(x_i, d)}{1 - \sum_i D_t(i) y_i h'(x_i, d)} \right).$$

  $\chi(a, d)$ shifts and normalizes a from $(0, 1)$ to $(-1, +1)$ and takes d as the new origin since $q_\lambda \in (0, 1)$.
- Choose $\alpha_t = \alpha'(d_t^*)$.
- Update $D_{t+1}(i) = \frac{D_t(i) \exp\{-\alpha_t y_i h_t(x_i)\}}{Z_t}$, where Z_t is the normalization factor to ensure that $D_{t+1}(\cdot)$ is a probability distribution.

The final hypothesis for any input sample x is as follows.

$$H(x) = \text{sign} \left(\sum_{t=1}^{T} \alpha_t h_t(x) \right).$$

Embedding the ME model into the boosting approach is quite intuitive. During the parameter estimation and feature induction, we could take the weights $D_t(i)$ as the sample priors $p(x_i)$ in [1], where they are originally assumed to be uniform distributions.

3.2.3. Decision threshold maximizing the $F1$ measure (BST-ME-F1)

In [5], the authors suggest choosing the weak classifier h_t at each iteration to minimize the exponential loss Z_e, as shown in Equation 2. This will facilitate the overall minimization of training errors. However, in some detection problems, different types of errors may be combined in specific ways. For example, a popular metric used in information retrieval applications is $F1$ - a balanced way of combining precision and recall (see the footnote on page 1). In this variation of boosting, we replace Z_e-minimization with $F1$-maximization. In the following section, we will find such simple replacement does slightly improve the precision-recall curve in news story segmentation.

3.3. SVM approaches

SVM has been shown to be a powerful technique for discriminative learning [3]. It focuses on structural risk minimization by maximizing the decision margin. We applied SVMs using the Radial Basis Function (RBF) as the kernel,

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0.$$

In the training process, it is crucial to find the right parameters C (tradeoff on non-separable samples) and γ in RBF. We apply five fold cross validation with a grid search with varying (C, γ) on the training set to find the best parameters achieving the highest accuracy. We conduct SVMs for story boundary detection on three feature sets, including (1) *35-dimension ME-induced binary features (SVM-BIN-35)*: we want to evaluate the synergy of the ME induced features with the SVM model (with optimal parameters $C = 2^{8.0}$ and $\gamma = 2^{-8.0}$); (2) *all binary features with 195 dimensions (SVM-BIN-195)* (with optimal parameters $C = 2^{1.7}$ and $\gamma = 2^{-4.6}$); (3) *33-dimension raw multi-modal features (SVM-RAW-33)*: we use the original raw multi-modal features without binarization and delta operations (with optimal parameters $C = 2^{2.2}$ and $\gamma = 2^{-5.8}$).

4. EXPERIMENTS

To train a CNN boundary detection model with A+V modalities, we use 34 CNN videos ($\sim$17 hours) with 1142 annotated boundaries and 11705 candidate points. Each candidate is with 195 binary features. The performance is measured in a separate test set with 22 CNN videos with totally 795 story boundaries and are represented in precision vs. recall curves shown in Figure 2. Breakdowns of errors made by different methods over different types of video stories (as defined in Figure 1) are shown in Table 1.

4.1. Performance of boosting approaches

From the results in Figure 2, performance of a single ME with 35 binary features (ME-BIN-35) is still the upper bound for those from the ensemble approaches, where BST-ME-F1-6x35 performs almost the same as ME-BIN-35. BST-BIN-35 is the 35-iteration boosting on binary features and takes binary features directly as weak learners. It performs the worst in the whole ensemble variants. This is probably because of the limited strengths of each single binary feature in predicting the result. Linear combinations of

such weak classifiers through boosting still cannot pull the performance to the comparable level. Moreover, an ME model selects the features jointly achieving the highest gain and usually skips features with redundant capabilities. Instead, the boosting approach might pick up the same feature repeatedly over the iterations, and thus might miss the larger gain achievable by using multiple features jointly. It is interesting to note in Table 1, BST-BIN-35 can only achieve comparable performance for video story types with simple syntax; e.g. types (a), (g), and (h).

The results in Figure 2 also confirm the performance gained by replacing the exponential loss criterion Z_e minimization with the $F1$ measure maximization. As discussed earlier, this is partly due to the direct relation of F1 metric to the precision-recall criteria.

According to the above experiment, the boosting approaches do not further improve the performance by using ME alone. It might be that the remaining (weighted) samples are still difficult after the early iterations of boosting. Thus, those weighted samples could not be solved by using different features, or current available features are not enough to discriminate them well.

4.2. Performance of SVM approaches

An SVM with all 195-dimension binary features performs the best among all the methods compared. The discriminative power of SVM takes advantages of the high dimension features generated from the feature wrapper without any reduction. SVM-BIN-35 is an SVM with 35-dimension binary features induced from the ME approach and performs slightly better than ME-BIN-35. Interestingly, it also performs better than SVM-RAW-33 in high F1 region (when precision and recall rates are similar). This seems to indicate that the feature subset selected by ME induction are indeed an effective set when detecting video story boundaries.

With the same binary feature set, SVM-BIN-35 performs better than ME-BIN-35 almost for every story type, shown in Table 1. It might imply that SVMs provide more discrimination power based on the principal of structure risk minimization or margin maximization on the decision boundary. Generative models like ME attempt to approximate the distribution of the input space. However, a good approximation of feature distribution may not lead to superior discrimination accuracy.

It is interesting to identify the challenging types of story boundaries from results shown in Table 1. Even with the best performance, SVM-BIN-195, there are still high miss rates for types (b), (d), (e), and (f), where types (d) and (e) are both large groups constituting more than 36% of the whole data set when combined. Most experiment setups perform well on types (a) and (c) which are dominated by the anchor face and prosody features. Understanding gained by such detailed analysis of error sources will be very useful for developing enhanced techniques in the future work.

5. CONCLUSION AND FUTURE WORK

Story segmentation in news video remains a challenging issue due to the diversity of production rules and feature dynamics. We believe that multi-modality fusion through effective statistical modelling and feature selection are keys to solutions. In this paper, we investigate different learning algorithms based on generative, discriminative, and ensemble models for multi-modal fusion and provide systematic performance analysis.

Specifically, we explore extension and combination of ME, Boosting, and SVM based approaches. We conduct extensive experiments over the TRECVID 2003 data set (CNN channel). Results indicate SVM-based methods are more effective in terms of classification accuracy while ME-based induction methods are effective in finding salient feature sets. The analysis of error sources over different types of video data also points out critical avenues for focusing future research efforts. We believe that the algorithmic approaches and experimental methodologies adopted in the study are powerful and will be valuable for other statistical detection tasks such as event detection and structure mining.

One of our future directions is to explore the temporal dynamics of the news program since the statistical behaviors of features in relation to the story transition dynamics may change over time in the course of a news program.

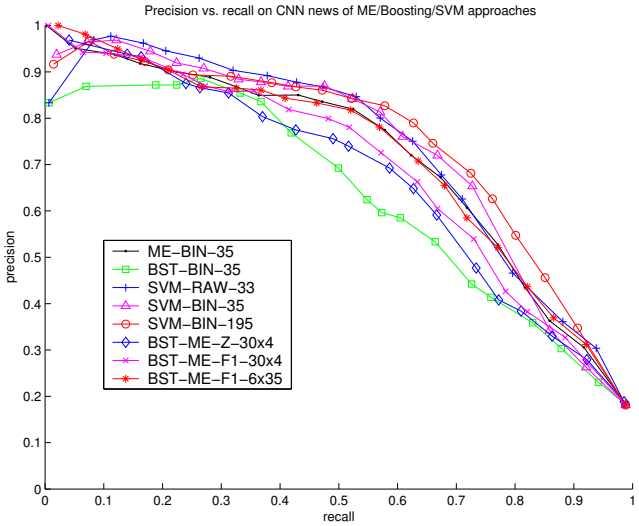

Fig. 2. Precision vs. recall curves of ME/boosting/SVM approaches. The definitions of legends are in Section 3, except that "BST-ME-(Z or F1)-$t\times n$" " is a BoostME combined method using Z_e-minimization or $F1$-maximization, t iterations, and n features in each induction step.

6. REFERENCES

[1] W. Hsu, S.-F. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *IS&T/SPIE Electronic Imaging*, San Jose, CA, 2004.

[2] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, no. special issue on Natural Language Learning, pp. 177–210, 1999.

[3] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[4] Y. Freund and . E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995.

[5] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, 1999.

[6] K. Tieu and P. Viola, "Boosting image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.