EFFECTIVE REPRESENTATIONS FOR LEVERAGING LANGUAGE CONTENT IN MULTIMEDIA EVENT DETECTION

Shuang Wu, Xiaodan Zhuang, Pradeep Natarajan

Speech, Language and Multimedia Business Unit,
Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA 02138
{swu,xzhuang,pradeepn}@bbn.com

ABSTRACT

Language content in videos from speech and overlaid or in-scene video text can provide high precision signals for video event detection and retrieval. However, sporadic occurrence, content that is unrelated to the events of interest, and high error rates of current speech and text recognition systems on consumer domain video make it difficult to exploit these channels. In this paper, we study different representations of language content to address these challenges. First, we utilize likelihood weighted word lattices obtained from a Hidden Markov Model (HMM) based decoding engine to encode many alternate hypotheses, rather than relying on noisy single best hypotheses. Second, we utilize an event-independent modified term frequency-inverse document frequency (TF-IDF) weighting scheme to obtain the final feature vector. We present detailed experimental results on the TRECVID MED 2013 dataset containing ∼150000 videos, and show that our representation significantly outperforms alternate representations for both speech and video text.

Index Terms— multimedia event detection, speech recognition, video text OCR, lattices, TF-IDF

1. INTRODUCTION

The ability to search through large volumes of digital videos and summarize their content has important applications. These include multimedia event detection (MED) to retrieve videos containing a target event, and event recounting to produce succinct summaries that are indicative of events contained in the videos. These tasks correlate with the recent research emphasis on consumer domain videos, e.g., in the TRECVID evaluations [1]. Compared to similar applications on video data from professional sources like broadcast news, consumer video analysis brings extra challenges such as heterogeneous topics and genres, varying media quality, and less structured photography and editing. Most systems performing these tasks leverage multiple sub-systems operating on various modalities [2, 3, 4, 5, 6].

Language content analysis, such as automatic speech recognition (ASR) and video text recognition (OCR), provides important information in consumer video analysis for multiple reasons. First, sub-systems targeting language content can leverage strong prior information to offset the challenge of noisy data. For example, word n-gram statistics constrain predicted word sequences to be those more likely for a given language. Second, language content analysis generalizes better than low-level analysis [7, 8], enabling effective adaption of off-the-shelf models trained using external data in related but different domains. For example, we can apply systems trained on broadcast news or scanned document data on the consumer video domain. Third, ASR and OCR bring significant value particularly for semantic analysis of video, lending concise yet highly informative content for both downstream machine learning and human interpretation.

Spoken content analysis in consumer video presents many challenges. Consumer videos are highly heterogeneous due to variability in content, style, production qualities and language. Speech can be spontaneous, conversational and lack of inherent structure. Background noise and speech from multiple background speakers often overlay on the acoustic signal of the target speaker in consumer videos. Due to the large volume of such heterogeneous data, techniques relying on manually labeled data are also impractical. Last but not the least, a viable approach must be scalable and able to handle a large, or varying, set of target events in a huge archive, requiring minimum human intervention. Video text imposes similar challenges in this domain.

Given these challenges, language content analysis in consumer videos faces inevitable errors from even state-of-the-art ASR and OCR systems. This has been a typical challenge for spoken language processing systems [9]. Recognized words contain errors, and even correctly recognized words are not equally indicative of target multimedia events.

In this paper, we describe our system for language content analysis in consumer domain videos. On the audio chan-

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied , of IARPA, DoI/NBC, or the U.S. Government.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

nel, we first perform speech activity detection (SAD) to identify likely speech segments. On the visual channel, we apply video text detection to identify the bounding boxes of text lines. Both ASR and OCR are performed using Hidden Markov Model (HMM) based systems with multi-pass decoding, mainly trained on non-consumer video data. Both systems provide hypothesized language content in the form of likelihood weighted word lattices, encoding a rich yet compact summary of many alternative hypotheses.

We study various noise-robust vector representations based on these word lattices, enabling effective kernel support vector machine (SVM) classification. In particular, word counts are weighted with lattice arc posteriors. Different weighting and normalization methods are compared, including our novel modified TF-IDF based encoding, to determine the optimal setup. These vector representations are general enough to support the detection of many different high-level events without the need for event specific vocabularies. We present large scale experimental results on the TRECVID MED dataset [1], showing the superior performance of our representation for both ASR and OCR.

2. RELATED WORK

Many MED systems leverage both low-level and semantic components operating on audio and visual modalities [3, 4, 5, 6]. Existing works discuss ways of leveraging noisy language content in multimedia data mostly focus on ASR output [10, 11]. More word weighting schemes have been explored for text categorization and ranking [12, 13]. This paper studies both speech and video text content in consumer video, both encoded in the form of word lattices. By using two state-of-the-art HMM-based ASR and OCR systems, we demonstrate the correlation between effective methods applied to both sources of language content.

Weighting or pruning words using weights such as TF-IDF [14, 9] are common practices in natural language processing. However, recent literature [11] as well as our previous experiments find it non-trivial to improve performance using such common practices for the task of high-level event detection using error-prune ASR and OCR outputs. For example [11] shows that simply using the original ASR vocabulary with the logarithm of the expected word counts outperforms many different combinations of vocabulary pruning and weighting. Our previous work uses a modified TF-IDF, involving word frequencies in each event, to identify informative words [10]. However, the TRECVID 2013 MED task requires detecting events unknown at indexing time, therefore event-independent representations.

We demonstrate a unique word vector representation, constructed using statistics based on modified TF and IDF with no need for event definitions or positive samples. The proposed representation consistently improves high-level event detection performances for both ASR and OCR.

3. SPEECH PROCESSING

We use GMM-based speech activity detection (SAD) and HMM-based multi-pass large vocabulary automatic speech recognition (ASR) to obtain speech content in the video, and encode the hypotheses in the form of word lattices.

The SAD and ASR models, as well as their training and offline adaptation data, are described in more details in [10]. We provide the performances on the consumer video data for reference: SAD obtains a false alarm rate of 10.1% and missed detection of 5.8% according to the NIST *md-eval* metric, with a 0.25 second collar. The WER of the baseline ASR system is 48.2%, and the WER of the ASR system with offline adapted language model and dictionary is 35.8%. The system outputs not only the 1-best transcripts but also word lattices with acoustic and language model scores.

4. VIDEO TEXT PROCESSING

We use an HMM-based multi-pass large vocabulary OCR on text regions provided by a video text detector. Similar to the ASR, word lattices are used to encode alternative hypotheses. We leverage a statistically trained video text detector based on SVM to estimate bounding boxes. This detector is developed based on [15] with improved front-end processing based on Maximally Stable Extremal Regions (MSER).

On a small consumer video dataset with annotated video text bounding boxes, the detector achieves pixel-level precision and recall of 67.9% and 31.8% respectively. Note that these measurements are calculated on the raw pixel level, as our HMM-based OCR system expects tight bounding boxes around video text regions. The BBN HMM-based OCR systems is detailed in [16]. The video text content exists in various forms, such as subtitles, markup titles, and in-scene text (e.g., banners and road signs), and is therefore much more challenging than conventional scanned document OCR. Since we focus on bag-of-words representation for OCR in this work, we measure the word precision and recall within each video, at 37% recall and 14.7% precision.

5. LEVERAGING NOISY OUTPUT

5.1. Posteriors from word lattices

Given that the ASR and OCR systems above are processing very noisy data, the performance of 1-best transcripts is not satisfactory as shown in subsections 3 and 4. Further, among the words in the 1-best output, some may be of very low confidence. To alleviate the negative impact on event detection, we use word lattices or confusion networks [17] instead of the 1-best transcripts. This has been widely used for keyword spotting [18]. Compared to the 1-best transcripts, a lattice contains more alternatives and is more likely to recover important keywords missed by errors in the 1-best transcripts.

Words in the different arcs of the lattice L_v from video v have different confidences. We score word w using the arc posterior probabilities $\{p_a(w)\,|a \in L_v\}$ derived via the forward-backward algorithm [19]. Aggregated posterior probabilities of a word can serve as the expected count,

$$C(w) = \sum_{a \in L_v} p_a(w). \tag{1}$$

One can collect the expected counts from the lattices and create a word histogram to represent the audio data [18]. To avoid the noise created by words with very low posterior probability, we can use a threshold to remove word arcs with posterior probability lower than a threshold.

A histogram can also be constructed using only the 1-best in a similar fashion where every word in the 1-best is counted as one occurrence. We previously established the significant benefit of using expected word counts computed from lattice arc posteriors [10]. Therefore, we adopt this method, instead of using 1-best outputs, for experiments in this paper.

5.2. Effective representations with discriminative weights

In the interest of storage and processing efficiency as well as ability to generalize to diverse events in large video collections, we build a representation that is event independent. We start with a simple histogram representation built on the full lattice vocabulary after removing stopwords, which we denote W. This representation,

$$h_v = \{C(w)\,|\,w \in W\}\,, \tag{2}$$

is simply a histogram, i.e. the expected counts $C(w)$ for each word w for video v. The vector h_v is ℓ_1 normalized to produce a normalized histogram $\hat{h}_v$.

Since the lattice vocabulary is quite large and inclusive, we reduce the size of the representation by removing both extremely rare words in the vocabulary which are too infrequent to be robustly processed by machine learning techniques, as well as common words which are not useful in distinguishing videos of different events. Removing these extrema also help to offset noise from falsely detected speech and video text content. To this end, we compute the posterior-weighted document frequency $df(w)$ of each word $w \in W$ over a sample collection of videos V as

$$df(w) = \sum_{v \in V} \min\left(1, C(w)\right). \tag{3}$$

We then remove the 200 most common words as ranked by df, producing an upper cutoff $\overline{df}$. We also set a lower cutoff $\underline{df}$ by removing words with frequencies less than 0.001 times $\overline{df}$. These two cutoffs produce the compact vocabulary

$$W' = \left\{w \,\middle|\, w \in W,\, \underline{df} < df(w) < \overline{df}\right\}. \tag{4}$$

In our experiments, these cutoffs remove approximately $1/3$ of the words in W for both ASR and OCR.

Using the shortened vocabulary W', we have

$$f_v = \{C(w)\,|\,w \in W'\}\,. \tag{5}$$

Rather than ℓ_1 normalizing f_v, we normalize by the ℓ_1 of h_v, i.e., by the sum of posteriors of all words in the lattice of the given video, to produce $\tilde{f}_v$. This way, we encode in the normalization term information about the total amount of hypothesized language content in video v, to discount the potential false alarm in video text detection and speech detection.

We can further encode information about the potential classification power of each word in our vocabulary W' by weighting them by a revised inverse document frequency,

$$idf(w) = \log\left(\overline{df}/df(w)\right). \tag{6}$$

We then define

$$g_v = \{C(w) \times idf(w)\,|\,w \in W'\}\,. \tag{7}$$

We produce two representations by two different normalization schemes for g_v: first, with a simple ℓ_1 normalization to produce an idf-weighted histogram $\hat{g}_v$, and second, with the ℓ_1 norm of h_v, as we did with $\tilde{f}_v$, to produce $\tilde{g}_v$.

All the above computation can be performed on the fly for each single video with precomputed df and idf statistics on some sample videos. Our unique method of filtering and weighting the vocabulary by leveraging the lattice information from decoding, as well as incorporating video-specific information about the amount of hypothesized language content in the normalization, differentiate our representation from those weighting methods explored but shown ineffective in [11] for the same high-level event detection task.

6. MULTIMEDIA EVENT DETECTION EXPERIMENTS

6.1. MED setup

We test our approach on the TRECVID 2013 MED dataset [1], which is a large collection of ∼150000 consumer web videos containing 20 diverse high-level multimedia events. The dataset contains two prespecified training conditions for the events of interest: EK100, for which 100 positive exemplars of each event are provided along with a fixed set of ∼5000 background videos; and EK10, where only 10 positive exemplars are provided with the same background collection. We carry out multimedia event detection experiments using either audio (ASR) or visual (OCR) language information, and report results on the designated MEDTest set containing ∼25000 videos (including between 16 and 234 positive instances for each event). We use the provided Research set, which contains ∼12000 background videos and no exemplars of the events of interest, to compute the df and idf

Modality	Representation	EK100 (MAP)	EK10 (MAP)
ASR	$\hat{h}_v$	11.01%	3.93%
ASR	$\tilde{f}_v$	**12.04%**	**6.68%**
ASR	$\hat{g}_v$	10.28%	3.70%
ASR	$\tilde{g}_v$	**12.38%**	**6.00%**
OCR	$\hat{h}_v$	6.47%	2.16%
OCR	$\tilde{f}_v$	**8.26%**	**2.81%**
OCR	$\hat{g}_v$	6.75%	2.18%
OCR	$\tilde{g}_v$	**7.84%**	**3.25%**

Table 1. SVM-based event detection performance using different word vector representations on the MEDTest set.

statistics for our representation. All word processing is case-insensitive.

An SVM classifier [20] is trained for each event, using only the provided training collections. SVM hyper-parameters for misclassification cost and kernel width are estimated through extensive cross-validation grid search within the training set, for each representation and each event separately. Performance is measured on the MEDTest set by mean average precision (MAP) across the 20 events. In preliminary experiments, we found that the χ^2 kernel produces the most competitive results compared to linear and RBF kernels, and therefore use it in all experiments below.

6.2. Experimental results

For ASR and OCR, we perform the MED task using representations $\hat{h}_v$, $\tilde{f}_v$, $\hat{g}_v$, and $\tilde{g}_v$ presented in Section 5.2. From Table 1, we observe that $\tilde{f}$ and $\tilde{g}$ significantly outperform $\hat{h}$ and $\hat{g}$ in both EK100 and EK10 conditions. We believe that $\tilde{f}$ effectively leverages the aggregated posteriors of the larger vocabulary in representing total video language content, while reducing noisy decoding errors and redundant common words through the restricted vocabulary W'. Further, $\tilde{g}$ in some instances demonstrates the additional benefit of incorporating the $idf(w)$ term that additionally weights words by their relative frequencies on the research set. Note that $\hat{g}$ does not perform as well as either $\tilde{f}$ or $\tilde{g}$, indicating the importance of using the full speech content in computing the normalization term when reducing vocabulary size.

6.3. TRECVID results

Independent evaluations were conducted by NIST as part of the TRECVID 2013 evaluations on a blind ∼100000 video dataset, both for the same 20 events as in Section 6.2 (*pre-specified*), as well as for 10 new events given one week before the evaluation (*ad hoc*). Figure 1 summarizes our system performance compared to all other participating teams for which NIST released ASR and OCR results. Our ASR and OCR systems, based on the $\tilde{g}$ representation, achieved highly competitive scores across both EK100 and EK10 training conditions as well as prespecified and ad hoc events. In all conditions, our system is within the top two performing submissions. Note that our representation generalizes well to the ad hoc events with no vocabulary modification.

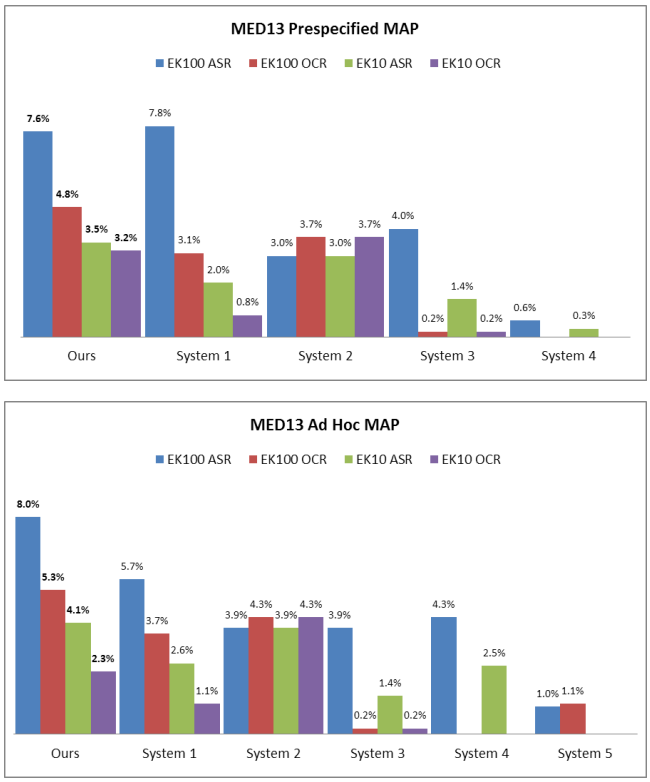

Fig. 1. TRECVID 2013 evaluation results for our ASR and OCR systems against all other ASR/OCR submissions.

7. CONCLUSION

To leverage speech and video text recognition in consumer video, we explore different representations of language content which are robust to noisy recognition output and are compact and effective for detection of diverse events. We utilize the likelihood weighted word lattices obtained from a Hidden Markov Model (HMM) based decoding engine to efficiently encode many alternate hypotheses, rather than relying on noisy single best hypotheses. We additionally leverage the information encoded in the lattices to generate a novel event-independent modified TF-IDF weighting using a two-tiered vocabulary to produce a compact and robust representation. We present detailed experimental results on the TRECVID MED 2013 dataset and show that despite previous works [11] indicating otherwise, our carefully constructed normalization and weighting scheme outperforms alternate representations for both speech and video text. Our results from the official TRECVID 2013 evaluation show that our representation generalizes well to unseen events of interest.

8. REFERENCES

[1] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot, "Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2013*. NIST, USA, 2013.

[2] Liangliang Cao, Shih-Fu Chang, Noel Codella, Courtenay Cotton, Dan Ellis, Leiguang Gong, Matthew Hill, Gang Hua, John Kender, Michele Merler, Yadong Mu, Apostol Natsev, and John R. Smith, "IBM Research and Columbia University TrecVid-2011 Multimedia Event Detection (MED) System," in *Proceedings of NIST TrecVid Workshop*, MD, 2011.

[3] Lei Bao, Longfei Zhang, Shoou-I Yu, Zhen zhong Lan, Lu Jiang, Arnold Overwijk, Qin Jin, Shohei Takahashi, Brian Langner, Yuanpeng Li, Michael Garbus, Susanne Burger, Florian Metze, and Alexander Hauptmann, "Informedia@TRECVID 2011," in *Proceedings of NIST TrecVid Workshop*, MD, United States, 2011.

[4] Pradeep Natarajan, Prem Natarajan, Vasant Manohar, Shuang Wu, Stavros Tsakalidis, Shiv N. Vitaladevuni, Xiaodan Zhuang, Rohit Prasad, Guangnan Ye, Dong Liu, I-Hong Jhuo, Shih-Fu Chang, Hamid Izadinia, Imran Saleemi, Mubarak Shah, Brandyn White, Tom Yeh, and Larry Davis, "BBN VISER TRECVID 2011 multimedia event detection system," in *Proceedings of NIST TrecVid 2011 Workshop*, Gaithersburg, MD., 12 2011.

[5] Cees G. M. Snoek, Koen E. A. van de Sande, Xirong Li, Masoud Mazloom, Yu-Gang Jiang, Dennis C. Koelma, and Arnold W. M. Smeulders, "The MediaMill TrecVid 2011 Semantic Video Search Engine," in *Proceedings of NIST TrecVid Workshop*, MD, United States, 2011.

[6] Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah, "High-level event recognition in unconstrained videos," *International Journal of Multimedia Information Retrieval*, pp. 1–29, 2012.

[7] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[8] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *CVPR*. 2010, pp. 3384–3391, IEEE.

[9] Xiaodong He and Li Deng, "Speech-centric information processing: An optimization-oriented approach," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1116–1135, 2013.

[10] Stavros Tsakalidis, Xiaodan Zhuang, Roger Hsiao, Shuang Wu, Pradeep Natarajan, Rohit Prasad, and Prem Natarajan, "Robust event detection from spoken content in consumer domain videos," in *INTERSPEECH'12*, 2012, pp. –1–1.

[11] Julien van Hout, Murat Akbacak, Diego Castan, Eric Yeh, and Michelle Sanchez, "Extracting spoken and acoustic concepts for multimedia event detection," in *ICASSP'13*, 2013, pp. –1–1.

[12] Amit Singhal, Chris Buckley, and Mandar Mitra, "Pivoted document length normalization," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 1996, SIGIR '96, pp. 21–29, ACM.

[13] Edda Leopold and Jrg Kindermann, "Text categorization with support vector machines. how to represent texts in input space?," *Machine Learning*, vol. 46, no. 1-3, pp. 423–444, 2002.

[14] Gerard Salton and Christopher Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988.

[15] Xujun Peng, Huaigu Cao, Rohit Prasad, and Premkumar Natarajan, "Text extraction from video using conditional random fields," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp. 1029–1033.

[16] Xujun Peng, Huaigu Cao, Srirangaraj Setlur, Venu Govindaraju, and Prem Natarajan, "Multilingual ocr research and applications: an overview," in *Proceedings of the 4th International Workshop on Multilingual OCR*, New York, NY, USA, 2013, MOCR '13, pp. 1:1–1:8, ACM.

[17] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, pp. 373–400, 2000.

[18] Shasha Xie and Yang Liu, "Using confusion networks for speech summarization," in *NAACL-HTL*, 2010.

[19] G. Evermann and P.C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *ICASSP*, 2000, vol. 3, pp. 1655–1658.

[20] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.