TEMPORAL LOCALIZATION OF AUDIO EVENTS FOR CONFLICT MONITORING IN SOCIAL MEDIA

Junwei Liang, Lu Jiang and Alexander Hauptmann

School of Computer Science, Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

With the explosion in the availability of user-generated videos documenting any conflicts and human rights abuses around the world, analysts and researchers increasingly find themselves overwhelmed with massive amounts of video data to acquire and analyze useful information. In this paper, we develop a temporal localization framework for intense audio events in videos which addresses the problem. The proposed method utilizes Localized Self-Paced Reranking (LSPaR) to refine the localization results. LSPaR utilizes samples from easy to noisier ones so that it can overcome the noisiness of the initial retrieval results from user-generated videos. We show our framework's efficacy on localizing intense audio event like gunshot, and further experiments also indicate that our methods can be generalized to localizing other audio events in noisy videos.

Index Terms— Gunshot detection, Audio event detection, Self-paced Learning, Audio reranking

1. INTRODUCTION

Tremendous amount of videos is being uploaded to social network sites every minute, documenting every aspect of our lives and events all over the world. This boom of ever-increasing video data has resulted in an explosion in the availability of documentation and visual evidence for any unexpected conflicts events or wrong-doing, such as the Boston Marathon Bombing in 2013 and the Shooting of Dallas police in 2016. Such incidents usually happen during a major event where a large crowd of people are gathered and recording the surroundings with personal devices. As audio-visual surveillance data are often unavailable at the scene, such huge volume of user-generated videos are probably the only source for researchers and analysts to investigate the situation. However, due to the daunting amount of videos, it is almost impossible for analysts to extract useful information manually in a short time to assist law enforcement for immediate action [1].

Automatic methods for detection and localization of intense events in a large video collection [1], such as gunshot, explosion, are important for analysts and researchers to uncover useful evidence that is widely dispersed over time or spread through different videos captured by multiple people. For example, such system can help analysts to pinpoint video segments in Dallas Shooting that contain gunshots so that they can locate the shooters and analyze the situation. People in these intense events are frightened or confused, and thus the captured visual information are often too blurry, obscured or unstable to be useful. On the contrary, audio becomes a reliable signal for detecting an event. For example, gunshots or explosion sound can be robustly detected in videos with very low resolution.

A major research challenge in audio event detection lies in temporal content localization which not only needs to detect which videos have the interested events, but also to temporally localize the events within the videos. To address this challenge, this paper proposes a novel method called Localized Self-Paced Reranking (LSPaR). Reranking has been an important technique in retrieval to improve initial search results, and has proven to be effective in a variety of problems [2]. The proposed LSPaR advances the state-of-the-art reranking method by allowing it to temporally localize audio events within a video clip. Since the initial ranked list is noisy, LSPaR trains a reranking model starting from easy samples with more confidence score from the top of the list and then gradually incorporates noisier samples later. Such easy to noisy strategy has proven to be efficient in the learning of noisy data [3]. To verify its efficacy and robustness, we conduct experiments on three datasets for localizing the gunshot event, where LSPaR significantly outperforms existing baseline methods. We also experiment on other common audio events and demonstrate that LSPaR can generalize to discover general audio event in videos.

2. RELATED WORK

Audio Event Detection: Many early works on soundtrack analysis focused on distinguishing a small number of sound classes such as speech, music, silence or noise. They were solved with various traditional machine learning and signal processing approaches [4, 5, 6]. Such works were mostly done on clean broadcast or television program audio date [5]. With the increasing number of user-generated videos available, many research works were done for this more challenging data. We can categorize soundtrack analysis work into sub-soundtrack classification or entire soundtrack classification. Recently there are a number of efforts to classify short video clips into a fixed number of sound classes. The TRECVID multimedia event detection (MED) annual evaluation organized by the NIST[1] is a representative benchmark [7]. The MED using soundtrack is the entire soundtrack classification problem[7]. Many works compared different acoustic features and discriminative models for modeling event [8, 9, 10]. Similar approaches were used for sub-soundtrack classification on user-generated videos [11].

There were also many works that specialized in gunshot detection. One focused on detecting gunshot in movies using dynamic programming and Bayesian networks [12]. Audio-Surveillance Systems were also used with various acoustic features for scream and gunshot detection [13]. A two-stage approach was used to improve efficiency and reliability of gunshot detection systems [14]. These approaches were specially designed for gunshot detection and may not be able to generalize to other audio event detection. In this paper, we show that our method not only produces accurate gunshot detection results but also demonstrates that it can be generalized to detect other audio events.

Reranking: Reranking methods were first used in text retrieval [15]. Most reranking methods are unsupervised methods that includes classification [16], clustering [17] and Learning-to-rank based reranking [18]. The TRECVID annual evaluation organized by NIST included a content-based search task, the multimedia event detection (MED) 0Ex (Zero-Example) [7]. Under such criteria, where only text queries were provided to perform a content-based video search, the reranking methods showed significant improvement over the plain retrieval result [2].

Recently Kumar et al. designed a learning paradigm, called *self-paced learning* (SPL) [19]. The learning theory is inspired by the underlying cognitive processes of humans and animals, which generally start with learning easier aspects of a task, and then gradually take more complex examples into consideration [19]. This theory has been successfully applied to various applications, such as action/event detection [20], weakly supervised learning from the Internet [3], tracking [21] and segmentation [22], reranking [2], etc. Following the idea, after initial retrieval results of possible segments that may include the audio event, our framework learns a reranking model iteratively from first using a few segments within the videos with more confident scores, then incorporates more noisy segments.

3. AUDIO EVENT LOCALIZATION FRAMEWORK

The proposed audio event temporal localization framework contains the following key components as shown in Figure 1.

Audio Feature Representation. In order to temporally localize audio event within a video, we first extract soundtrack from the videos and chunk the audio stream into small segments with overlap (exp. 3-sec window and 1-sec shift). We incorporate the widely used Bag-of-Words (BoW) features[8, 11] in our system. In this paper we apply this model to the low level MFCC features.

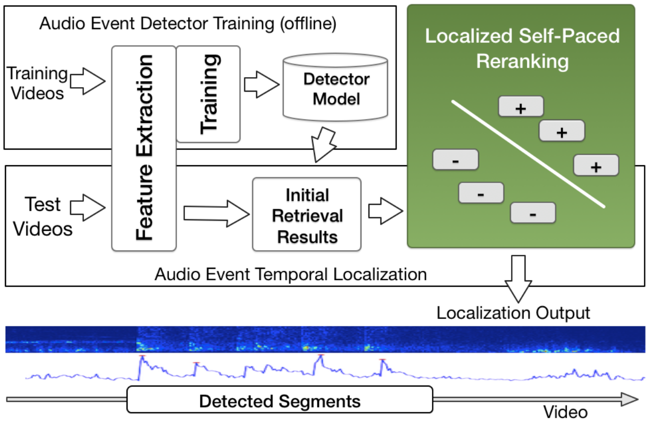
Audio Event Detector Training (offline)
Training Videos
Feature Extraction
Training
Detector Model
Test Videos
Initial Retrieval Results
Localized Self-Paced Reranking
+
+
+
-
-
-
Audio Event Temporal Localization
Localization Output
Detected Segments
Video

Fig. 1. **Audio Event Temporal Localization System**

Event Detector Model. After we extract audio features, we train two-class SVM classifiers for each audio events and then apply them to the video segments from test videos. However, because of the nature of the user-generated videos, the initial detection results have low accuracy due to noise.

Localized Self-Paced Reranking. For the test videos, after the detector model produces an initial ranked list of video segments, we utilize LSPaR to learn a reranking model iteratively from first using a few video segments with more confident scores at the top of the list, then incorporates more noisy segments. Detailed algorithm is described in the following section. Finally after smoothing we merge the segment results and output the final localization result.

4. LOCALIZED SELF-PACED RERANKING

4.1. Objective Function

To overcome the noise of the initial detection result, we propose Localized Self-Paced Reraning (LSPaR). Formally, given a set of $\mathcal{D}$ of n video segments, Let $L(\tilde{y}_i, g(\mathbf{x}_i, \mathbf{w}))$, or l_i for short, denote the loss function which calculates the cost between the pseudo label $\tilde{y}_i$ and the estimated label $g(\mathbf{x}_i, \mathbf{w})$. $\tilde{y}_i \in \{-1, 1\}$ is the pseudo label for the ith video segment whose value is assumed since the true labels are unknown. Here $\mathbf{w}$ represents the model parameter inside the decision function g. For example, in our paper, $\mathbf{w}$ represents the weight parameters in the Support Vector Machine (SVM). Our objective function is to jointly learn the model parameter $\mathbf{w}$, the pseudo label $\tilde{y}$ and the latent weight variable $\mathbf{v} = [v_1, \cdots, v_n]^T$ for the video segments by:

$$\min_{\mathbf{w},\mathbf{v},\tilde{y}} \mathbb{E}(\mathbf{w},\mathbf{v},\tilde{\mathbf{y}};\lambda,\Psi) = \sum_{i=1}^{n} v_i L(\tilde{y}_i, g(\mathbf{x}_i,\mathbf{w})) + f(\mathbf{v};\lambda,p), \quad \text{s.t. } \tilde{y}_i \in \{-1,1\}, \mathbf{v} \in \Psi \tag{1}$$

where L is the standard hinge loss, calculated from:

[1] https://www.nist.gov/

$$L(\tilde{y}_i,g(\mathbf{x}_i,\mathbf{w})) = \max\{0, 1-\tilde{y}_i g(\mathbf{x}_i,\mathbf{w})\} \tag{2}$$

$\mathbf{v} \in [0,1]^n$ denote the latent weight variables reflecting the pseudo labels' confidence, which determine a learning sequence for the video segments. Video segments with greater weights tend to be learned earlier. Our goal is to assign greater weights to the segments with more confident labels whereas smaller or zero weights to the segments with noisy labels. To this end, we employ the self-paced regularizer f, which controls the learning process of the model [20].

Ψ in Eq. (1) is a curriculum region [20] that incorporates the localization knowledge extracted from the initial detection result as a convex feasible region for the weight variables. The shape of the region indicates a prior learning sequence for the segments, where favored segments have greater expected values. The region is derived based on initial detection results, i.e. video segments in videos that have greater initial detection scores have greater expected values. For simplicity, in this paper, Ψ is only used in the initialization of $\mathbf{v}$.

We consider the linear regularizer Eq. (3) proposed in [20, 3] with dropout strategy to address the noise in the videos:

$$\begin{aligned} r_i(p) &\sim \text{Bernoulli}(p) + \epsilon, (0 < \epsilon \ll 1) \\ f(\mathbf{v};\lambda,p) &= \frac{1}{2}\lambda \sum_{i=1}^{n} (\frac{1}{r_i} v_i^2 - 2v_i), \end{aligned} \tag{3}$$

where $\mathbf{r}$ is a column vector of independent Bernoulli random variables with the probability p of being 1. The term dropout in this paper refers to dropping out samples in the iterative learning. By dropping out a sample, we drop out its update to the model. In practise, we only apply dropout to the selection of negative segments.

4.2. Algorithm

Algorithm 1: Localized Self-Paced Reranking.

```
input  : Input a initial ranked list of $\mathcal{D}$ and a step size $\mu$
output: Reranked list $\mathcal{D}^*$
1 Initialize $\mathbf{v}^*$, $\tilde{\mathbf{y}}^*$, $\lambda$, and $\Psi$ based on $\mathcal{D}$;
2 while not converged and not reach max iteration do
3     Update $\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbb{E}(\mathbf{w}, \mathbf{v}^*, \tilde{\mathbf{y}}^*; \lambda, \Psi)$;
4     Update $\tilde{\mathbf{y}}^* = \arg\min_{\tilde{\mathbf{y}}} \mathbb{E}(\mathbf{w}^*, \mathbf{v}^*, \tilde{\mathbf{y}}; \lambda, \Psi)$;
5     Update $\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathbb{E}(\mathbf{w}^*, \mathbf{v}, \tilde{\mathbf{y}}^*; \lambda, \Psi)$;
6     if $\lambda$ is small then increase $\lambda$ by the step size $\mu$;
7 end
8 Apply $\mathbf{w}$ to $\mathcal{D}$ to get $\mathcal{D}^*$;
9 return $\mathcal{D}^*$
```

Eq.(1) is difficult to optimize directly due to its non-convexity and complex constraints. Following [19, 2], we employ the alternative convex search algorithm to solve Eq. (1). The algorithm divides the variables into three blocks, i.e. classifier parameters $\mathbf{w}$, pseudo labels $\tilde{\mathbf{y}}$ and latent weight variables $\mathbf{v}$. Algorithm 1 takes the input of the initial ranked list and a step size parameter, and outputs the reranked list. First of all, it initializes the pseudo label and the latent weight variables in the feasible region based on the initial ranked list. Then it alternates among three steps until it finally converges: Step 3 learns the optimal model parameter with the most recent $\mathbf{v}^*$ and the pseudo labels $\tilde{\mathbf{y}}$. We use probabilistic sampling based on $\mathbf{v}$ to select samples for Liblinear [23] to train the model. In step 4 we learn the optimal pseudo labels. Since $\tilde{\mathbf{y}}$ is independent of $\mathbf{v}$ and $\tilde{y}_i \in \{-1, 1\}$, we can optimize $\tilde{\mathbf{y}}$ by:

$$\tilde{\mathbf{y}} = \arg\min_{\tilde{\mathbf{y}}} \sum_{i=1}^{n} L(\tilde{y}_i,g(\mathbf{x}_i,\mathbf{w})) \tag{4}$$

where $\tilde{\mathbf{y}}$ denotes the optimal pseudo label. We can find the optimal $\tilde{\mathbf{y}}$ by enumerating each $\tilde{y}_i$. Step 5 learns the optimal weight variables with the fixed $\mathbf{w}^*$ and $\tilde{\mathbf{y}}$. Substituting Eq. (3) into Eq. (1), we can optimize $\mathbf{v}^* = [v_1^*, \cdots, v_n^*]^T$ by:

$$v_i^* = \begin{cases} r_i(-\frac{1}{\lambda}\ell_i + 1) & \ell_i < \lambda \\ 0 & \ell_i \geq \lambda \end{cases}. \tag{5}$$

The underlying intuition of the self-paced learning can be inferred from the solution in Eq. (5). When a video segment's loss based on the current optimal pseudo label is less than the model "age" λ, it will be assigned a soft weight for learning in the next iteration, otherwise it will not be selected. Meanwhile, the dropout strategy also affects the selections of the segments. The model "age" is gradually increased so that more noisy segments will be incorporated in the training of a "mature" reranking model. In fact, Algorithm (1) is optimizing an underlying non-convex robust loss on the noisy data, which tends to depress samples with noisy labels or outliers. It theoretically justifies the efficacy of the proposed method [24].

5. EXPERIMENTS

In this section, we empirically verify the efficacy and robustness of our proposed method on two tasks. Since in most conflict or violence situation, the gunshot event is often the interested audio event, we apply our method to temporally localizing gunshot in videos. In the second task, we perform experiments on localizing a wider range of audio event and show that our method can be generalized. The code and datasets can be downloaded from our project website[2].

Given the initial ranked lists of video segments or the reranked ones, we use the average precision (AP) to evaluate the event localization performance for audio event:

$$AP = \frac{1}{r} \sum_{j=1}^{n} I_j \times \frac{r_j}{j} \tag{6}$$

where r is the total number of relevant segments of that event, n as the total amount of segments, I_j=1 when the jth segment is relevant otherwise I_j=0. A segment is considered relevant if the target event occurs in over half of that segment. r_j is

[2] http://www.cs.cmu.edu/~ junweil/audioLocalization

Table 1. Gunshot Temporal Localization Experiments

Baselines	Conflict Videos	TREC	Urban
Without Reranking	0.322	0.223	0.912
CPRF	0.369	0.226	0.916
Learning to Rank	0.316	0.218	0.893
MMPRF	0.379	0.229	0.916
LSPaR	**0.424**	**0.267**	**0.927**

the number of relevant segments in the first j segments. In the case of localizing multiple events, we report mean average precision (MAP) as our evaluation metric.

Baselines: The proposed method is compared against the following four baseline methods which cover both the classical and recent representative reranking algorithms. 1) *Without Reranking* is the initial retrieval method without reranking. 2) *CPRF* is a classification-based reranking method. Following [25], SVM classifiers are trained using the top-ranked video segments and bottom-ranked ones in the initial ranked list. 3) *Learning to Rank* is a ranking method mainly used in web queries. A LambdaMART [26] in the RankLib toolkit is used to train the RankSVM model. 4) *MMPRF* is a method that the reranking model is trained through t iterations with top-ranked and bottom-ranked segments [27], in which $m = 1$ and $t = 3$.

Our model: Algorithm 1 is used to solve Eq. (1) and the max iteration is set to 3. For efficiency, we incorporate Explicit feature mapping [28] with χ^2 kernel in all methods.

5.1. Gunshot Temporal Localization

Training: Since there is no existing large dataset specifically for gunshot detection in real-life noisy user-generated videos, we collect about 2,200 weakly labeled gunshot segments, totally over 100 minutes, from freesound.com. We also collect over 4,000 gunshot segments from Youtube.com, totaling 4 hours of data. For negative training samples, we collect dozens hours of everyday life videos.

Testing: To verify the efficacy of the proposed method, we test our system on 3 datasets from 3 different domains. 1) *Real-life Conflict Scene Videos (Conflict Videos):* We collect 10 real-life videos taken by unprofessional people during the shooting of Dallas police[3] and the shooting in Nigeria[4]. The longest video for localizing gunshot is 44 minutes long with only about 2 minutes of gunshot segments. 2) *TRECVID Gunshot Videos (TREC):* We collect 57 videos from TRECVID SIN task [7] that contain gunshots. 3) *UrbanSound Gunshot Videos (Urban):* We collect 117 audios that contain gunshot from the UrbanSound dataset [29]. We perform experiments on these three datasets and report the average precision on each dataset in Table 1.

In Table 1, the best results are highlighted. The proposed method has achieved a relative improvement of 31.7%, 19.7%, 1.6% on three datasets, respectively. For localizing gunshot in Urbansound dataset, since the original retrieval result is already very accurate, the improvement of LSPaR is minimal. In Real-life Conflict Videos Dataset, the proposed method significantly outperforms other baselines. As LSPaR learns from easy to hard samples, it is able to overcome the noisy environmental sound in conflict videos and thus more suitable to be used in such kind of data. A fully functional gunshot temporal localization application is released [5], where users can upload any video for online gunshot localization.

Table 2. Experiments of Audio Events on 10-Split Urban

Baselines	Precision@5	Precision@10	MAP
Without Reranking	0.540 ± 0.002	0.517 ± 0.002	0.566 ± 0.001
CPRF	0.543 ± 0.003	0.520 ± 0.001	0.579 ± 0.002
Learning to Rank	0.527 ± 0.004	0.514 ± 0.002	0.569 ± 0.002
MMPRF	0.544 ± 0.003	0.522 ± 0.002	0.581 ± 0.001
LSPaR	**0.578 ± 0.007**	**0.554 ± 0.005**	**0.609 ± 0.003**

5.2. Experiments on Generic Audio Events

In this experiment, we try to verify LSPaR's performance on localizing a wider range of audio events. We conduct our experiment on the full UrbanSound dataset to temporally localize 9 audio events other than gunshot in 1302 audios. We randomly split the dataset into 70/30 for training and testing 10 times to reduce the bias brought by the partition. We train our audio detectors for each audio event in the train split and test them on the test set. The reranking methods are applied on the initial testing ranked lists on the test set. Mean precision at top 5, top 10 results and mean average precision of all the audio events are reported. In Table 2 we show the mean and 95% confidence interval of each metric. As we see, the proposed method also significantly outperforms other baselines on 9 other audio events on average. For the event "air_condition", LSPaR improves initial results by a relative 20.8% (from 0.516 to 0.623 absolute AP). The experiment results again substantiate the rationality and robustness of the proposed method. The promising results on other audio events verify that the proposed method can be generalized.

6. CONCLUSION

In this paper, we proposed a novel reranking framework for temporal localization of audio events in noisy consumer videos for conflict and violence monitoring in social media. LSPaR extracts informative information from the noisy initial ranked list and improves the retrieval results by a significant margin. The result suggests LSPaR can not only work on intense audio events like gunshot, but also can be generalized to other kinds of audio events like "street_music" and "children_playing".

[3] https://en.wikipedia.org/wiki/2016_shooting_of_Dallas_police_officers

[4] https://www.youtube.com/watch?v=rRCBZXRoVBo

[5] http://aladdin1.inf.cs.cmu.edu/daisy/index.php/application/cGunshot, username and password are both "demo".

7. REFERENCES

[1] Jay D Aronson, Shicheng Xu, and Alex Hauptmann, “Video analytics for conflict monitoring and human rights documentation technical report,” 2015.

[2] Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann, “Easy samples first: Self-paced reranking for zero-example multimedia search,” in *MM*, 2014.

[3] Junwei Liang, Lu Jiang, Deyu Meng, and Alexander Hauptmann, “Learning to detect concepts from webly-labeled video data,” .

[4] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaten, “Content-based classification, search, and retrieval of audio,” *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[5] John Saunders, “Real-time discrimination of broadcast speech/music,” .

[6] Eric Scheirer and Malcolm Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. IEEE, 1997, vol. 2, pp. 1331–1334.

[7] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quenot, Maria Eskevich, Robin Aly, and Roeland Ordelman, “Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking,” in *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[8] Qin Jin, Peter F Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze, “Event-based video retrieval using audio,” in *Proceedings of INTERSPEECH*, 2012, p. 2085.

[9] Keansub Lee and Daniel PW Ellis, “Audio-based semantic concept classification for consumer video,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1406–1416, 2010.

[10] Anurag Kumar, Pranay Dighe, Rita Singh, Sourish Chaudhuri, and Bhiksha Raj, “Audio event detection from acoustic unit occurrence patterns.,” .

[11] Junwei Liang, Qin Jin, Xixi He, Gang Yang, Jieping Xu, and Xirong Li, “Detecting semantic concepts in consumer videos using audio,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2279–2283.

[12] Aggelos Pikrakis, Theodoros Giannakopoulos, and Sergios Theodoridis, “Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 21–24.

[13] Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, Fabio Antonacci, and Augusto Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. IEEE, 2007, pp. 21–26.

[14] Talal Ahmed, Momin Uppal, and Abubakr Muhammad, “Improving efficiency and reliability of gunshot detection systems,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 513–517.

[15] Linjun Yang and Alan Hanjalic, “Supervised reranking for web image search,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 183–192.

[16] Rong Yan, Alexander Hauptmann, and Rong Jin, “Multimedia search with pseudo-relevance feedback,” in *International Conference on Image and Video Retrieval*. Springer, 2003, pp. 238–247.

[17] Winston H Hsu, Lyndon S Kennedy, and Shih-Fu Chang, “Video search reranking via information bottleneck principle,” in *Proceedings of the 14th ACM international conference on multimedia*. ACM, 2006, pp. 35–44.

[18] David Grangier and Samy Bengio, “A discriminative kernel-based approach to rank images from text queries,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 8, pp. 1371–1384, 2008.

[19] M Pawan Kumar, Benjamin Packer, and Daphne Koller, “Self-paced learning for latent variable models,” in *NIPS*, 2010.

[20] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann, “Self-paced curriculum learning,” 2015.

[21] James Steven Supancic and Deva Ramanan, “Self-paced learning for long-term tracking,” in *CVPR*, 2013.

[22] M Pawan Kumar, Haithem Turki, Dan Preston, and Daphne Koller, “Learning specific-class segmentation from diverse data,” in *ICCV*, 2011.

[23] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, “Liblinear: A library for large linear classification,” *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[24] Deyu Meng and Qian Zhao, “What objective does self-paced learning indeed optimize?,” *arXiv preprint arXiv:1511.06049*, 2015.

[25] Alexander G Hauptmann, Michael G Christel, and Rong Yan, “Video retrieval based on semantic concepts,” *Proceedings of the IEEE*, vol. 96, no. 4, pp. 602–622, 2008.

[26] Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao, “Adapting boosting for information retrieval measures,” *Information Retrieval*, vol. 13, no. 3, pp. 254–270, 2010.

[27] Lu Jiang, Teruko Mitamura, Shoou-I Yu, and Alexander G Hauptmann, “Zero-example event search using multimodal pseudo relevance feedback,” in *International Conference on Multimedia Retrieval*. ACM, 2014, p. 297.

[28] Andrea Vedaldi and Andrew Zisserman, “Efficient additive kernels via explicit feature maps,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 3, pp. 480–492, 2012.

[29] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, “A dataset and taxonomy for urban sound research,” in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.