LATE FUSION AND CALIBRATION FOR MULTIMEDIA EVENT DETECTION USING FEW EXAMPLES

Julien van Hout[1], Eric Yeh[1], Dennis C. Koelma[2], Cees G.M. Snoek[2]
Chen Sun[3], Ramakant Nevatia[3], Julie Wong[1], Gregory K. Myers[1]

[1]SRI International, Menlo Park, USA
[2]University of Amsterdam, The Netherlands
[3]University of Southern California, Los Angeles, USA

ABSTRACT

The state-of-the-art in example-based multimedia event detection (MED) rests on heterogeneous classifiers whose scores are typically combined in a late-fusion scheme. Recent studies on this topic have failed to reach a clear consensus as to whether machine learning techniques can outperform rule-based fusion schemes with varying amount of training data. In this paper, we present two parametric approaches to late fusion: a normalization scheme for arithmetic mean fusion (*logistic averaging*) and a fusion scheme based on logistic regression, and compare them to widely used rule-based fusion schemes. We also describe how logistic regression can be used to calibrate the fused detection scores to predict an optimal threshold given a detection prior and costs on errors. We discuss the advantages and shortcomings of each approach when the amount of positives available for training varies from 10 positives (10Ex) to 100 positives (100Ex). Experiments were run using video data from the NIST TRECVID MED 2013 evaluation and results were reported in terms of a ranking metric: the mean average precision (mAP) and $\overline{R}_0$, a cost-based metric introduced in TRECVID MED 2013.

Index Terms— multimedia event detection, late fusion, score calibration, score normalization, system fusion

1. INTRODUCTION

As the quantity of online user-submitted multimedia content grows, indexing and reliably searching for specific content becomes increasingly challenging. Moreover, the data is very heterogeneous and often of poor audio or visual quality, which challenges the accuracy of current event detection technologies. The track of multimedia event detection conducted under the TRECVID evaluations by NIST aims to solve the problem of detecting specific events like "changing a tire" or "grooming an animal" in a heterogeneous corpus of user-submitted video clips. Accurately detecting such precise events requires input from various analysis channels (image and motion analysis, audio concepts, speech content, character recognition, etc.) that we will refer to as MED modalities. The best-performing approaches to solving this task use various modalities and combine their detection scores in a late-fusion scheme. It is to be noted that while some researchers have successfully developed early-fusion schemes [1, 2] to combine different modalities together and learn joint classifiers, not all modalities can be combined in this way and these systems still rely heavily on late-fusion as a final combination stage [2].

Approaches to late-fusion mainly fall into two categories: rule-based or statistical model based. Simple rule-based fusions (like arithmetic mean or geometric mean) which first normalize the scores to a comparable range and then treats each modality identically are popular for their inherent robustness to over-fitting [3, 4, 5]. Other rule-based techniques (like weighted averaging) use different weights for different modalities. Here, the weights are found using grid-search [6, 7], are set to a measure of the performance of each modality [3], or to a measure of confidence of the score [8]. Machine learning alternatives like logistic regression [5, 9], ridge regression [7], linear support vector machine (SVM) [9] and explicit optimization of an evaluation metric [9] have also been explored. While the above studies often compare multiple fusion techniques to one another, their conclusions can vary widely. For instance, [9, 6] claim gains from logistic regression fusion compared to arithmetic mean or grid-based search techniques while other studies found the opposite conclusion on a similar task [3, 4, 5]. Given the diversity of the modalities to fuse across research teams, and the fact that their scores show very different distributions (Gaussian, exponential, bimodal, etc.), we believe that such conflictive conclusions could be explained by differences in the modalities' score distributions, differences in the type of features used in learning-based techniques, and differences in the way missing values are handled. Unfortunately, these details are often overlooked in the above studies, making it difficult to draw definite conclusions.

The challenge of handling missing values in late fusion is very common in detection tasks, especially in MED where modalities' scores can go missing for various reasons: no audio was available, no speech was detected, no motion was detected in the video, etc. Traditional ways of dealing with missing features in late-fusion include inferring the missing scores from the mean of scores from other videos or setting the missing score to be the minimum score. Ideally, one would like to not make any assumption about the missing score's value but rather learn its value for various events and modalities. Such an approach has been successfully applied to other detection tasks such as speaker identification [10] or spoken term detection [11] by using a logistic regression framework with binary side-information.

While MED performance is usually measured in terms of mean average precision (mAP), we also considered the R_0 metric

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. Disclaimer: 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.

introduced in TRECVID MED 2013. R_0 can be interpreted as a risk based on costs of misses and false alarms that the system should minimize by picking the right threshold. The main challenge when optimizing such a metric is to properly calibrate the scores such that a good threshold can be chosen. Also, a fusion strategy that gave the best mAP might not be optimal in terms of R_0, since the two metrics target different use cases. Prior work in speaker detection [12] has found logistic regression to be a very efficient approach to both calibration and fusion over a wide range of operating points.

In this paper, we will introduce a late-fusion framework based on logistic regression, that handles missing features as binary side-information. We also introduce a novel discriminative normalization scheme for arithmetic mean called *logistic averaging* that is robust to limited number of training examples. Finally, we present a strategy to calibrate the final scores and pick optimal thresholds for R_0 and report MED results for both the mAP and R_0 metrics.

2. DESCRIPTION OF MODALITIES

In this section, we describe the scope of our individual modalities and how they were trained. A more in depth description of each modality can be found in [13]

Low-level visual features We extracted low-level visual features for two frames per second from each video. We followed the bag-of-codes approach, which considers spatial sampling of points of interest, visual description of those points, and encoding of the descriptors into visual codes. We used a mixture of SIFT, TSIFT, and C-SIFT descriptors [14]. We computed the descriptors around points obtained from dense sampling and reduced them to 80 dimensions with principal component analysis. We encoded the color descriptors using Fisher vectors with a Gaussian Mixture Model codebook of 256 elements [15].

Semantic visual features We detected semantic concepts for each frame using low-level visual features and following the approach in [16]. We trained 1,346 concept detectors based on linear SVMs. Each frame is then represented by the concatenated detector scores from all these concepts.

Visual event classifiers We included three visual event classifiers based on low-level and semantic features. To arrive at a video-level representation for the low-level visual event classifier, we relied on simple frame averaging. For the two video event classifiers based on semantic features, we aggregated the concept vectors per frame into a video-level representation. One approach used averaging and normalization, while the other method used semantic encoding. On top of both concept representations, we used an SVM with χ^2 kernel.

Low-level motion features The two low-level motion features were based on Dense Trajectories (DTs) [17] and MoSIFT [18]. We computed DT raw features with a step size of 10 pixels and MoSIFT raw features with default parameters. The raw features were encoded using first- and second-order Fisher Vector descriptors with a two-level spatial pyramid [19]. Descriptors were aggregated across each video. We generated four event classifiers: two with DT features using first- and second-order Fisher Vector descriptors, and two with MoSIFT features using first- and second-order Fisher Vector descriptors. A Gaussian-kernel SVM was used for classification, and the outputs from the same low-level feature were averaged.

Motion event classifiers Two event classifiers were generated based on action concept detectors. There were 96 action concepts annotated on the MED11 Event Kit provided by Sarnoff/UCF, and 101 action concepts from UCF 101 [20]. The action concept detectors were applied to small segments of videos and encoded by Hidden Markov Model Fisher Vector descriptors [21]. SVM with Gaussian kernel was used to train two event classifiers, one for each set of action concepts.

Low-level audio content For our audio features, we extracted mel-frequency cepstral coefficients (MFCCs) over a 10-ms window. MFCCs describe the spectral shape of audio. The first and second derivatives of the MFCCs were also computed. The MFCC features were difference-coded with Fisher vectors using a 1024-element Gaussian Mixture Model and classified using a linear SVM.

Spoken content We ran an English ASR model trained on conversational telephone data and adapted to meeting data. We performed supervised acoustic model adaptation using in-domain annotated TRECVID data and unsupervised adaptation using the first-pass recognition output. We also performed supervised and unsupervised language model adaptation. The lattice-based approach described in [22] was used to build the MED classifier, and the final score was the distance to the hyperplane of a L1-regularized linear SVM model, mapped to $[0, 1]$ by using a logistic function.

Written content SRI's English video OCR software detected and recognized text appearing in the TRECVID MED 2013 video imagery. This software recognizes both overlay text, such as captions that appear on broadcast news programs, and in-scene text on signs or vehicles [23]. For each event, we generated event profiles from the event descriptions by using term frequency-inverse document frequency (TF-IDF) weightings to rank the relevance of non-stop-words. The event detection score for each video was the cosine similarity between the word vector for the video and the word vector for the event profile.

3. LATE FUSION

In this section, we present approaches to late fusion that were used in our experiments. The scores x_i from each of the N modalities are detection probabilities and therefore lie in $[0, 1]$. The goal of late-fusion is estimating the probability of the label y of a given video given the score vector $\mathbf{x} = [x_0, x_1, ..., x_{N-1}]$.

3.1. Baseline fusions

We describe a few simple, widely used baseline fusion methods.

Arithmetic mean This method combines scores from various modalities by taking the arithmetic mean of the scores for each trial. We considered two ways of dealing with missing scores. In the *AM-zero* technique, a missing score is supposed to have a zero value. In the *AM-mean* technique, a missing score is supposed to have the mean value of the non-missing scores from other modalities. The latter technique is equivalent to computing the average over the non-missing scores only.

Geometric mean This method, referred to as *GM*, computes the fused score for a given trial as the geometric mean of all non-missing and non-zero scores for that trial.

Weighted averaging In this technique, the final score is computed as a weighted sum of the scores for each modality. The weights are chosen by optimizing the mAP metric through an exhaustive grid-search with weights taking values in $\{0.001, 0.01, 0.03, 0.1, 0.3, 0.6, 1\}$. The weights are trained on the cross-validation scores on the training data, and applied to the test data. We study two different setups with varying number of trained

parameters: in *WM-dep* the weights are event-dependent and in *WM-indep* the weights are optimized for all 20 events at once.

3.2. Logistic regression fusion

Logistic regression is a common approach for converting a M-dimensional vector of scores into a single value, the likelihood ratio, which can be used to make binary decisions. The LR model assumes that the posterior of a certain clip being a positive has the form $P(y = +1|\mathbf{x}) = \sigma(\alpha\mathbf{x}^T + \beta)$ where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. The parameters $\alpha = [\alpha_0, ..., \alpha_{M-1}]$ and β are learned by maximizing the L_2-regularized likelihood of the model on labeled training data by using the "Trusted Region Newton Method" [24] as implemented in the scikit-learn library [25]. The regularization parameter was tuned using cross-validation.

We propose to apply logistic regression to MED late fusion, a technique we refer to as *LR*, as follows. For each trial, we create a feature vector by concatenating the logit of scores of all of the N modalities, where the logit function is defined by $logit(x) = log(x/(1-x))$. The logit expands the dynamic range of the exponentially distributed probabilistic scores. The resulting scores are close to normally distributed for both positives and negatives and behave better for logistic regression. Missing scores are set to zero, and a feature is added for each modality as a binary indicator variable I_{miss} accounting for the possibility of missing scores for some trials. Initially introduced in [11] for late fusion of keyword spotting systems, this approach is equivalent to learning a bias for the missing score value of each modality. Once parameters are trained, the final posterior is given by:

$$P(y = +1|\mathbf{x}) = \sigma(\sum_{i=0}^{N-1} (\alpha_{2i}x_i + \alpha_{2i+1}I_{miss}(x_i)) + \beta)$$

Additionally, we propose to automatically perform feature selection and discard some modalities during training by looking at the trained weights α_{2i}. If the weight corresponding to a certain modality is found to be negative, the logistic regression is retrained with that modality removed. This approach is based on the intuition that a negative weight indicates an anti-correlation between the score of some modality and the label of a video clip, which is the sign of a poorly performing modality. By discarding that modality, we reduce the noise in the data as well as the dimension of the feature vector and obtain better generalization properties. This approach will be referred to as *LR+fs*. We also considered the *LR-min* and *LR-min+fs* systems where a missing score is set to the minimum score of that modality on the training data. These two systems will provide a comparison point against the proposed missing-values handling scheme.

3.3. Logistic averaging

The logistic averaging technique, or *LA*, is a novel technique that non-linearly normalizes the scores of various modalities before performing arithmetic mean fusion.

As in the case of logistic regression, we apply the logit function to the posterior scores of our modalities to map them from $[0, 1]$ to $[-\inf, +\inf]$. We apply Z-normalization by computing the means and variances of the cross-validated logit-scores for each event on the training data. The same normalization is applied to the videos in the test set. Then we map those scores back to $[0, 1]$ using the sigmoid function $\sigma_{\alpha,\beta}$ defined as $\sigma_{\alpha,\beta} = 1/(1 + e^{-(\alpha x+\beta)})$. The parameters α and β are chosen to optimize the mAP on the cross-validated training scores using a grid search. If X_i denotes the Z-normalized logit-score from each modality, then the fused score is given by:

$$P(y = +1|\mathbf{x}) = \frac{1}{N}\sum_{i=0}^{N-1} \sigma_{\alpha,\beta}(X_i)$$

As mentioned in previous work on late fusion of biometric systems [26], Z-normalization performs best when the input distributions are Gaussian distributed. By applying the logit to our initial scores, which are exponentially distributed, we obtain near-Gaussian distributed scores. The role of α and β is to enable some non-linearity in the arithmetic mean fusion by tuning a sigmoid that modifies the modalities' score distributions. Because the logit-scores are normalized around 0 with variance 1, a small α would lead to a nearly linear mapping, while a large α introduces a sharp cutoff at $-\beta$, below which the scores are set to 0, and above which the scores are set to 1.

Though this approach does not train different weights for each modality and can therefore seem sub-optimal compared to weighted averaging or logistic regression, it is less prone to over-fitting as it does not rely on labeled positives to estimate the mean and variances. It does require some positives to tune α and β but as these parameters are fixed for all events, their estimation is quite robust.

4. CALIBRATION AND THRESHOLD SELECTION

Besides maximizing average precision, a second challenge of the MED 2013 TRECVID evaluation [27] is to select, for each event, the detection threshold t that maximizes the R_0 metric defined as: $R_0(t) = Rec(t) - \frac{K}{V}Rank(t)$ where $K = 12.5$, V is the total number of clips in the test set, $Rec(t)$ is the recall at threshold t and $Rank(t)$ is the number of clips whose score is larger than t. It can be shown that $R_0(t)$ can be rewritten as:

$$R_0(t) = C_1\left(C_2 - [(K\pi_+^{test})^{-1} - 1]N_{miss}(t) - N_{fa}(t)\right)$$

where C_1 and C_2 are constants, $N_{miss}(t)$ and $N_{fa}(t)$ are the respective number of misses and false alarms at threshold t, and π_+^{test} is the ratio of positives in the test set. The threshold that maximizes this quantity also minimizes the risk given by:

$$Risk(t) = C_{miss} \cdot N_{miss}(t) + C_{fa} \cdot N_{fa}(t)$$

where $C_{fa} = 1$ and $C_{miss} = (K\pi_+^{test})^{-1} - 1$. Bayesian decision theory indicates that in order to minimize this risk the system should decide that the clip with scores $\mathbf{x}$ is a positive if and only if

$$p(y = +1 \mid \mathbf{x}) \cdot C_{miss} > P(y = -1 \mid \mathbf{x}) \cdot C_{fa}$$

which defines a threshold on the log-likelihood ratio (LLR):

$$LLR = log\left(\frac{p(\mathbf{x} \mid y = +1)}{p(\mathbf{x} \mid y = -1)}\right) > log\left(\frac{C_{fa}}{C_{miss}}\right) - logit(\pi_+^{test})$$

This formulation comes in handy when using logistic regression to fuse or calibrate scores. Indeed, it can be shown that with a posterior of the form $P(y = +1|\mathbf{x}) = \sigma(\alpha\mathbf{x}^T + \beta)$, the following holds for the LLR:

$$LLR + logit(\pi_+^{train}) = \alpha\mathbf{x}^T + \beta$$

where π_+^{train} is the ratio of positives in the training set. Assuming that the scores $S = \alpha\mathbf{x}^T + \beta$ at the output of the logistic regression are well calibrated, the threshold t_0 that maximizes R_0 is therefore:

$$\begin{aligned} t_0 &= log\left(\frac{C_{fa}}{C_{miss}}\right) - logit(\pi_+^{test}) + logit(\pi_+^{train}) \\ &= logit(K\pi_+^{test}) - logit(\pi_+^{test}) + logit(\pi_+^{train}) \end{aligned}$$

It is worth noting that while π_+^{train} is known from the training data labels, π_+^{test} might not be known and the difference between the assumed and the actual π_+^{test} may result in a sub-optimal threshold.

5. EXPERIMENTAL RESULTS

In this section, we first describe the data used for our experiments and then present results on system fusion using two separate metrics.

5.1. Data

We evaluated the performance of late fusion according to the NIST TRECVID 2013 MED evaluation plan [27]. We used the 20 pre-specified events as our detection targets. We ran experiments in two conditions with varying numbers of positives: *100Ex* with 100 positive clips per event and *10Ex* with 10 positive clips. An extra set of 4992 video clips labeled as negatives is used to supplement the positives for each event. To maximize the use of this limited amount of training data, we generated scores on the training data for each modality using 10-fold cross-validation. These cross-validation scores were used to train all of the normalization and fusion parameters, as well as to choose thresholds. The MED performance is reported on MEDTEST, a set of 23,468 video clips labeled as negatives plus 1,489 video clips labeled as positives, for an average of 75 labeled positives per event.

5.2. Results and discussion

For logistic regression and weighted averaging, we reduced the number of trained parameters by merging together similar modalities using arithmetic mean in the posterior domain prior to learning the fusion. Specifically, we fused together all three visual modalities and all four motion-based modalities to create two aggregate modalities. For *10Ex* for logistic regression, we even averaged those two modalities into a single modality.

Columns 2 and 5 of Table 1 present the mean Average Precision (mAP) of individual sub-systems and of several fusion schemes on the test set. In both training conditions, the logistic regression and logistic averaging techniques significantly outperform the baseline fusion techniques. These gains can be explained by the greater flexibility of parametric approaches such as *LR*, *LA* and *WM*. Yet, the two proposed approaches do not suffer the over-fitting problems of *WM*, either because the number of trained parameters is small (*LA*) or because of regularization (*LR*). When 100 examples are available for training, the best technique is *LR+fs* with a mAP of 0.434. This results demonstrate the efficiency of using binary indicators to handle missing values since *LR-min+fs* obtained a mAP of 0.422 only. Also, while logistic regression proves more efficient than *LA* for *100Ex*, this trend disappears in the *10Ex* condition where both *LA* and *LR-min+fs* perform best. We believe that this result is directly related to the design of *LA* as a fusion technique not requiring many positives for training and tuned to optimize mAP. In contrast, logistic regression learns event-dependent weights that enable it to perform well over a wide range of operating points. The feature-selection component of logistic regression was found useful and provided a significant mAP increase in both conditions. Modalities were discarded for 7 events for *10Ex* and 16 events for *100Ex*.

We also compared the performance of the different fusion strategies in terms of the R_0 metric. We considered two different strategies to pick the threshold: (1) using the threshold t_{tr} that optimizes $R_0(t)$ on the training data, and (2) using the threshold t_0 computed as in the theoretical analysis developed in Section 4. The latter technique assumes that the fused scores are calibrated likelihood ratios. Because this is only the case for the LR fusion, we first calibrate the output of the other fusion strategies for each event by using a pass of logistic regression with a 2-dimensional feature vector. The two dimensions are set to the logit of the fused score in the posterior domain and a binary indicator that is set to 1 if the score is missing. The logistic regression parameters are estimated using the fused cross-validation scores on the training set. The α and β parameters of the logistic averaging approach were adjusted to maximize $R_0(t_0)$ on the training data.

Results in terms of $\overline{R}_0$, the mean of R_0 over all 20 events, are shown in Table 1. With this metric, logistic regression consistently outperforms the other fusion approaches, for both conditions. Logistic averaging remains competitive but no longer outperforms *LR-min+fs* in the *10Ex* condition. An advantage of logistic regression fusion over other techniques for maximizing R_0 is that it uses a feature vector with at least N dimensions, and therefore more finely approximates the LLR. The resulting scores are better calibrated than the scores obtained through a pass of logistic regression following late-fusion. Also, we should point out that the t_0 threshold obtained through Bayesian decision theory consistently outperforms the more ad-hoc technique of picking t_{tr} using the training data. This is because the Bayesian formulation enables computing the theoretically optimal threshold given the event detection priors on the test data.

Table 1: Performance of various score fusion techniques on the test set. Results are reported in both conditions in terms of mAP and $\overline{R}_0$. The best system for each metric is reported in bold.

System	100Ex			10Ex		
	mAP	$\overline{R}_0(t_{tr})$	$\overline{R}_0(t_0)$	mAP	$\overline{R}_0(t_{tr})$	$\overline{R}_0(t_0)$
AM-zero	0.405	0.514	0.524	0.222	0.263	0.278
AM-mean	0.402	0.506	0.526	0.214	0.244	0.272
GM	0.356	0.483	0.487	0.206	0.148	0.038
WM-dep	0.404	0.511	0.532	0.197	0.259	0.297
WM-indep	0.414	0.503	0.522	0.213	0.288	0.298
LA	0.425	0.521	0.531	**0.244**	0.287	0.304
LR-min	0.421	0.519	0.538	0.235	0.260	0.309
LR-min+*fs*	0.422	0.522	0.541	**0.244**	0.288	**0.314**
LR	0.428	0.532	0.537	0.230	0.262	0.312
LR+*fs*	**0.434**	0.533	**0.546**	0.238	0.295	0.312

6. CONCLUSION AND FUTURE WORK

In this paper, we demonstrate the efficiency of parametric approaches to late fusion in a multimedia event detection system, even in situations with limited training data. The proposed logistic regression approach to score fusion handles missing scores and automatically performs feature selection to discard poorly performing modalities. A second technique, proposed under the name of logistic averaging, can be seen as a pre-processing approach to the arithmetic mean method by performing Z-normalization in the logit domain before mapping scores back to posteriors in a way that maximizes a given metric. The logistic regression approach significantly outperformed baseline techniques in terms of both the mAP and the $\overline{R}_0$ metric. Logistic averaging was very competitive for optimizing mAP with limited training data, but didn't perform as well on the $\overline{R}_0$ metric. These findings are comparable to results for other detection tasks such as speaker identification or keyword spotting where logistic regression has consistently been found to be a robust tool to combine systems and provide a calibrated output that can be used to make binary decisions over a wide range of operating points. Avenues for future work include applying the fusion techniques introduced in this paper to the problem of query-based event detection, where the event detection models are built from an event description rather than learnt using positive examples.

7. REFERENCES

[1] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[2] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann, "Double fusion for multimedia event detection," in *Advances in Multimedia Modeling*, pp. 173–185. Springer, 2012.

[3] Gregory K Myers, Ramesh Nallapati, Julien van Hout, Stephanie Pancoast, Ramakant Nevatia, Chen Sun, Amirhossein Habibian, Dennis C Koelma, Koen EA van de Sande, Arnold WM Smeulders, et al., "Evaluating multimedia features and fusion for example-based event detection," *Machine Vision and Applications*, pp. 1–16, 2013.

[4] Arash Vahdat, Kevin Cannons, Hossein Hajimirsadeghi, Greg Mori, Scott McCloskey, Ben Miller, Sharath Venkatesha, Pedro Davalos, Pradipto Das, Chenliang Xu, et al., "Trecvid 2012 genie," in *Proceedings of 2012 TRECVID workshop*. NIST.

[5] Lei Bao, Shoou-I Yu, Zhen-zhong Lan, Arnold Overwijk, Qin Jin, Brian Langner, Michael Garbus, Susanne Burger, Florian Metze, and Alexander Hauptmann, "Informedia@ trecvid 2011," *TRECVID2011, NIST*, 2011.

[6] D Oneata, M Douze, J Revaud, J Schwenninger, D Potapov, H Wang, Z Harchaoui, Jakob Verbeek, Cordelia Schmid, R Aly, et al., "Axes at trecvid 2012: Kis, ins, and med," 2012.

[7] Liangliang Cao, N Codella, L Gong, et al., "Ibm research and columbia university trecvid-2012 multimedia event detection (med), multimedia event recounting (mer), and semantic indexing (sin) systems," in *Proc. TRECVID 2012 workshop. Gaithersburg, MD, USA*, 2012.

[8] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, and Rohit Prasad, "Multimodal feature fusion for robust event detection in web videos," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1298–1305.

[9] Ilseo Kim, Sangmin Oh, Byungki Byun, AG Amitha Perera, and Chin-Hui Lee, "Explicit performance metric optimization for fusion-based video retrieval," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 395–405.

[10] Luciana Ferrer, Lukas Burget, Oldrich Plchot, and Nicolas Scheffer, "A unified approach for audio characterization and its application to speaker recognition," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[11] Murat Akbacak, Lukas Burget, Wen Wang, and Julien van Hout, "Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams," in *Proceedings of ICASSP 2013*. IEEE.

[12] Niko Brümmer and Johan du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.

[13] Gregory K Myers, Ramesh Nallapati, Julien van Hout, Stephanie Pancoast, Ramakant Nevatia, Chen Sun, Amirhossein Habibian, Dennis C Koelma, Koen EA van de Sande, Arnold WM Smeulders, et al., "The 2013 sesame multimedia event detection and recounting system," *Proceedings of TRECVID 2013, to appear*.

[14] Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1582–1596, 2010.

[15] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, pp. 1–24, 2013.

[16] Amirhossein Habibian, Koen EA van de Sande, and Cees GM Snoek, "Recommendations for video event recognition using concept vocabularies," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 89–96.

[17] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[18] Ming-yu Chen and Alexander Hauptmann, "Mosift: Recognizing human actions in surveillance videos," 2009.

[19] Chen Sun and Ram Nevatia, "Large-scale web video event classification by use of fisher vectors," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 15–22.

[20] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[21] Chen Sun and Ram Nevatia, "Active: Activity concept transitions in video event classification," in *International Conference on Computer Vision (ICCV)*, 2013.

[22] Julien van Hout, Murat Akbacak, Diego Castan, Eric Yeh, and Michelle Sanchez, "Extracting spoken and acoustic concepts for multimedia event detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3657–3661.

[23] Gregory K Myers, Robert C Bolles, Quang-Tuan Luong, James A Herson, and Hrishikesh B Aradhye, "Rectification and recognition of text in 3-d scenes," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2-3, pp. 147–158, 2005.

[24] Chih-Jen Lin, Ruby C Weng, and S Sathiya Keerthi, "Trust region newton method for logistic regression," *The Journal of Machine Learning Research*, vol. 9, pp. 627–650, 2008.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] Anil Jain, Karthik Nandakumar, and Arun Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[27] NIST, "Trecvid multimedia event detection evaluation plan," 2013.