Explicit Performance Metric Optimization for Fusion-based Video Retrieval

Ilseo Kim[†], Sangmin Oh[‡], Byungki Byun[*], A. G. Amitha Perera[‡], Chin-Hui Lee[†]

[†]Georgia Institute of Technology, [‡]Kitware Inc., [*]Microsoft

Abstract. We present a learning framework for fusion-based video retrieval system, which explicitly optimizes given performance metrics. Real-world computer vision systems serve sophisticated user needs, and domain-specific performance metrics are used to monitor the success of such systems. However, the conventional approach for learning under such circumstances is to blindly minimize standard error rates and hope the targeted performance metrics improve, which is clearly suboptimal. In this work, a novel scheme to directly optimize such targeted performance metrics during learning is developed and presented. Our experimental results on two large consumer video archives are promising and showcase the benefits of the proposed approach.

1 Introduction

In many computer vision problems, the success of the learning algorithms is measured by domain- and application-specific performance metrics that simulate the real-world needs. One example is video retrieval, where diverse performance metrics are used to measure the quality of the system as well as the potential user experience. For example, [1,2] uses precision of top ranked retrieval results; TRECVID multimedia event detection (MED) task [3] prefers the ratio of 12.5:1 between probability of miss and false alarm; and [4] uses F-1 score. However, most learning methods optimize error rate, not the domain-specific performance measure, potentially yielding suboptimal solutions.

Another imperative aspect of real-world computer vision systems is the ability to fuse multiple features. The benefits of fusion have been clearly demonstrated in recent literature. For example, for video retrieval, [2,4] use multiple audio visual cues, and [1,5,4] incorporate even text features from tags or the video webpages. However, most of these techniques use the traditional hinge loss error function during their learning process and have not attempted to directly optimize their preferred performance metrics.

In this work[1][2], we propose a learning framework which is able to directly optimize specific performance metrics, and demonstrate its value in effectively

[1] This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright 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.

[2] We thank Anthony Hoogs, Greg Mori, Arash Vahdat, Zhi Feng Huang, Weilong Yang, Scott McCloskey, and Ben Miller for helpful discussion and sharing features used in this work. The third author contributed to this work while he was with Georgia Institute of Technology.

fusing multiple features. First, we introduce a systematic learning framework to *directly* optimize specific performance metrics beyond simple error rates. This direct optimization means that we can avoid the prolonged parameter search process typically required when the target metrics are optimized indirectly. Second, we apply our learning framework to fusion classifiers for consumer video retrieval problems. We show that our approach can learn competitive fusion classifiers while simultaneously optimizing given performance metrics. Our experiments on challenging video datasets show promising results.

2 Related Work

With our focus on optimizing performance metrics for fusion classifiers in consumer video retrieval, there are three areas of related work.

Performance Metric Optimization. Learning with explicit performance metric optimization has been mostly studied in machine learning community, albeit sparsely. [6] is a good reference and discusses optimization of a few performance metrics for SVM and boosting. However, most of them use elements of discrete search, different from our straightforward continuous optimization. Pareto criteria was introduced for multiple performance metric optimization [7]; however, Pareto criteria only provides partial ordering between multiple metrics, and joint optimization or complex metrics are not supported. In contrast, the basis of our approach, maximal figure-of-merit (MFoM), is a general framework which has been applied to problems such as text categorization, e.g., [8]. This work provides the first study on incorporating MFoM framework for audio-visual fusion for video retrieval, and presents the first principled approach to explicitly optimize the criteria in Sec. 3.

Fusion. An example fusion method is multiple kernel learning (MKL) [9]. In MKL, because a final fusion classifier is trained using all features jointly from early stages, it can be categorized as an 'early fusion' method. However, MKL does not systematically support the optimization of particular performance metrics, and reported results are not always competitive [10]. Other examples include the use of boosting for fusion [11,10]. A variant of LP-Boost introduced in [10] is more related to our work in terms of overall 'late fusion' architecture.

Fusion-based Consumer Video Retrieval. The fusion of multi-modal features for consumer video retrieval is an on-going area of research. For example, [2] introduced CCV dataset and showcase a benchmark system which uses SVM as a fusion classifier. [4] introduces a retrieval system which improves performance by incorporating manually designed semantic hierarchy. [5,1] presents tag recommendation approaches on YouTube videos. For collaborative competition, TRECVID [3] runs an MED track and disseminates large datasets annually.

3 Explicit Performance Metric Optimization

In this section, we show how our approach explicitly optimizes targeted performance metrics. The novel elements are in the details of incorporating the performance metric into the objective function of a learning framework. For clarity, this paper focuses on our chosen metric; however, the derivation can be easily extended to other metrics of interest.

3.1 Evaluation metric for real-world video retrieval

In real-world retrieval tasks, the performance metrics that capture user desires can differ widely. For example, for a 'Google search', the important metric may be precision of the top-N. For a statistical analysis problem, on the other hand, recall may be the most important factor. In general, a large class of these metrics can be thought of as the weighted combinations of the probabilities of missed detections (P_{MD}) and false alarms (P_{FA}) at a particular operating point.

In this paper, we focus on the weighted sum of P_{MD} and P_{FA} at a particular ratio, wihch is suggested by the TRECVID MED tasks. Concretely, the goal is:

$$\text{Minimize } S_\tau = P_{\text{MD}} + \tau \times P_{\text{FA}} \;\; \text{s.t.} \;\; \frac{P_{\text{MD}}}{P_{\text{FA}}} = \tau. \tag{1}$$

In the following, we explain our approach with regards to this particular metric. However, we note again that the framework is more general, and can be easily applied to other metrics such as rankings, F_1 or average precision.

To optimize the metric in Eq. 1, a standard scheme is to learn a model with its own learning objectives and adjust detection thresholds until the desired ratio of $P_{\text{MD}}/P_{\text{FA}} = \tau$ is met where the metric S_τ will be computed. With this approach, however, there is no guarantee that the learning procedure will focus on improving performance at particular operating points. Our solution described in the following sections provides a principled approach to achieve such a goal.

3.2 Maximal-Figure-of-Merit (MFoM) Framework

Our learning task is formulated within a discriminative framework. Let $T = \big\{(x, y) \,|\, x \in R^D,\, y \in C\big\}$ be a set of training data, where x is a D-dimensional sample and y is a class label $C = \{C_+, C_-\}$, i.e., positive and negative.

Let $d(x; \Lambda) \in (-\infty, \infty)$ be a *class confidence function* which indicates the confidence that a sample x belongs to the positive class, C_+, where a large positive value corresponds to a high confidence. Given $d(\cdot)$ and Λ, the decision rule for a sample x is defined as *accept* $x \in C_+$ if $d\,(x; \Lambda) > 0$, and *reject* otherwise. Our goal is to learn the parameters Λ to optimize the targeted metric.

The core ideas of our MFoM-based learning approach are two-fold. First, we exploit the fact that most performance metrics and their sub-components, such as P_{MD} and P_{FA} in Eq. 1, can be expressed as a combination of the four sub-metrics from a confusion matrix: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Second, we approximate a target metric such as S_τ in Eq. 1, that is based on discrete error counts with a parameterized continuous and differentiable loss function $L\,(T; \Lambda)$.

In particular, the four sub-metrics are approximated as continuous functions using (truncated) sigmoid functions $\sigma\,(\cdot)$, which approaches one for high confidence for a positive class C_+, or approaches zero otherwise. In detail, the four approximated sub-metrics are expressed as follows:

$$\widehat{\text{TP}} = \sum_{(x,y)|y\in C_+} \sigma\big(d(x;\Lambda)\big)\,, \quad \widehat{\text{FN}} = \sum_{(x,y)|y\in C_+} \big(1 - \sigma\big(d(x;\Lambda)\big)\big) \tag{2}$$

$$\widehat{\text{FP}} = \sum_{(x,y)|y\in C_-} \sigma\big(d(x;\Lambda)\big)\,, \quad \widehat{\text{TN}} = \sum_{(x,y)|y\in C_-} \big(1 - \sigma\big(d(x;\Lambda)\big)\big)$$

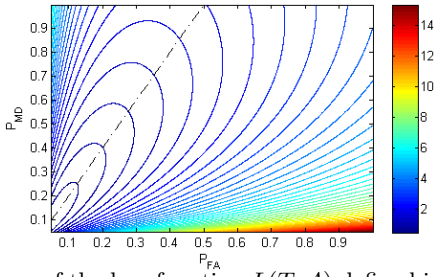

Fig. 1. Iso-contour curves of the loss function $L(T;\Lambda)$ defined in Eq. 4 when $\tau = 2$ and $\gamma = 1$. The dashed straight line corresponds to a iso-ratio $P_{\mathrm{MD}}/P_{\mathrm{FA}} = 2$.

where the sigmoid function, $\sigma(z) = \left(1 + \exp(-\alpha \cdot z)\right)^{-1}$, is parameterized by a positive constant α. Then, the overall loss function L is formulated from approximate sub-metrics ($\widehat{\cdot}$) using a mapping function $f(\cdot)$ as follows:

$$S_\tau \approx L\left(T;\Lambda\right) = f\left(\widehat{\mathrm{TP}}, \widehat{\mathrm{FP}}, \widehat{\mathrm{TN}}, \widehat{\mathrm{FN}} | \Lambda\right) \tag{3}$$

The role of the mapping function f is to reconstruct the loss function L accurately from sub-metrics. In fact, if the given target metric is a simple combination of sub-metrics, a precise mapping f is possible; e.g., for the F_1 metric where $F_1 = 2\mathrm{TP}/\left(2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}\right)$. In some cases, however, the loss function may involve complex conditions such as the ratio constraint in Eq. 1, which needs approximation. We discuss this issue further in Sec. 3.3.

Finally, the optimal parameter Λ_{opt} that minimizes $L\left(T;\Lambda\right)$ is learned by the generalized probabilistic descent (GPD) [12] algorithms.

In all, there are three steps needed for the MFoM framework to be properly used for problems at hand. First, an appropriate parameterized class-confidence function $d(x;\Lambda)$ needs to be defined. The class of linear discriminant functions (LDF) is used in this work; but, in general, any parameterized function can be used [8] such as a kernelized discriminant function. Second, a good mapping function f needs to be designed to simulate the target metric. Finally, an effective constant α which controls the slope of the sigmoid function needs to be selected. The larger α is, the more accurate the approximations in Eq. 2. However, the smaller α is, the smoother the overall approximation in Eq. 3. In practice, we observed that the choice of α affects convergence speed, rather than accuracy, for most datasets with reasonable sizes.

3.3 Strategies for Complex Target Metric Approximation

In this section, we present how a good mapping function f in Eq. 3 can be designed to yield an accurate continous loss function $L(T;\Lambda)$ for a given target metric, with focus on the example metric introduced in Eq. 1.

For cases where complex target metrics prohibit the use of precise mapping function f, our proposed method is to approximate the target metric as a combination of simpler sub-functions. This usually involves a set of parameters Γ which control the relative weights of sub-functions. Optimal values for Γ may be found through analytic approaches by minimizing the divergence between the resulting approximation f and the given target metric. On the other hand, good values for Γ can be found through cross validation as well. In fact, a more

complex scheme of dynamically varying Γ during learning can be beneficial. For example, in Eq. 1, an optimal value for Γ may differ according to varying values of P_{MD} and P_{FA} during learning steps. The investigation of diverse detailed learning strategies is beyond the scope of this work, so we focus on illustrating these ideas on a concrete example below.

For the example target metric in Eq. 3, a linear sub-function for weighted error rate $\left[\widehat{P_{\mathrm{MD}}} + \tau \times \widehat{P_{\mathrm{FA}}}\right]$ can be incorporated in a straightforward manner where the approximations $\widehat{P_{\mathrm{MD}}}$ and $\widehat{P_{\mathrm{FA}}}$ are set to be equal to $\widehat{\mathrm{FN}}$ and $\widehat{\mathrm{FP}}$ (in Eq. 2) divided by the total number of positive and negative samples respectively. In addition, our mapping function should be designed to prefer user-specified target ratio τ between P_{MD} and P_{FA}. To enforce such a ratio constraint, we include a sub-function $R\left(\tau,\, P_{\mathrm{MD}}/P_{\mathrm{FA}}\right)$ which monotonically increases loss with respect to the difference between a target ratio τ and the exhibited ratio $\widehat{P_{\mathrm{MD}}}/\widehat{P_{\mathrm{FA}}}$. By incorporating both terms with a weighting parameter $\Gamma = \gamma$, the loss function $L(T; \Lambda)$ that approximates Eq. 1 is finally defined as:

$$L\left(T; \Lambda\right) = \left[\widehat{P_{\mathrm{MD}}} + \tau \times \widehat{P_{\mathrm{FA}}}\right] + \gamma \times \left[R\left(\tau,\, \widehat{P_{\mathrm{MD}}}/\widehat{P_{\mathrm{FA}}}\right)\right] \tag{4}$$

With small γ, learning focuses more on minimizing the error rate; however, the learned model is less likely to show a desired target error ratio τ, since the minimum value of the weighted error rate could be derived by reducing P_{FA} and sacrificing P_{MD}, especially when τ is large. On the other hand, with large γ, learning will focus more on meeting target error ratio, and less on decreasing error rates. In this work, we set γ to a fixed constant by searching through cross-validation; this has shown promising results.

Among many options for the ratio constraint approximation term R, we found the following form to work well and used it in this work:

$$R\left(\tau,\, \widehat{P_{\mathrm{MD}}}/\widehat{P_{\mathrm{FA}}}\right) = \left\{\log\left(\tau\right) - \log\left(\frac{\widehat{P_{\mathrm{MD}}}}{\widehat{P_{\mathrm{FA}}}}\right)\right\}^2 \tag{5}$$

The logarithmic squared form used above provides a computational advantage in that overall gradients can be easily computed as a sum of two terms (i.e., the gradients of P_{MD} and P_{FA}), avoiding the complications potentially caused by the direct use of division $P_{\mathrm{MD}}/P_{\mathrm{FA}}$.

To showcase the quality of the approximation in Eq. 4, Fig. 1 illustrates the iso-contour curves of the loss function, along with the iso-error ratio line (dashed). It can be clearly seen that the designed loss function is correlated with and declines towards the iso-error ratio line. This implies that the minimum value of the loss function defined in Eq. 4 can be found near the iso-error ratio line and left-bottom of the plot through the gradient descent procedures.

4 Late Score Fusion Framework

Our fusion-based video retrieval architecture is formulated within the *late fusion* paradigm. By late fusion, we mean that scores are computed independently by multiple base classifiers, one per feature type, and fusion classification is conducted on the computed scores. We use the MFoM approach to learn the fusion classifier parameters while explicitly optimizing target performance metrics.

4.1 Training Discriminative Score Fusion

During training, each base classifier is trained in a one-vs-all manner as well, and is used to generate a single score for the target class. For base classifiers, we used SVMs and their estimated probabilities as base classifier scores.

For a fusion classifier, we used MFoM learning scheme and adopted LDF as our class-confidence function as $d(x;\Lambda) = \sum_j \omega_j x_j + \omega_0$, where x is the score vector from base classifiers. Accordingly, MFoM systematically learns the weights for each score dimension for the target class, while explicitly optimizing the desired performance metric. This way, the fusion classifier becomes confident when multiple base classifier scores are high, and vice versa.

4.2 Additional Non-Target Class Scores for Fusion Classifiers

To improve the performance of the fusion classifier further, we have investigated the use of additional non-target base classifier scores as inputs for 1-vs-All fusion classifiers, and observed consistent improvement in the final fusion classification. For example, we can incorporate the output by a base classifier trained for *Birthday party* for the training of a fusion system for the target class of *Wedding*. In this scheme, our fusion classifier uses $(M \times K)$-dimensional discriminative scores as its inputs, where there are K features and M base classifiers available. We believe the improvement is obtained because a fusion classifier systematically incorporates the correlation among event classes. Negative correlation as well as positive correlation could be helpful to acquire more discriminant power, i.e., high probabilities of outdoor event classes infer low confidence on indoor event classes. Fig. 5 illustrates the learned model parameters of fusion classifiers for the 10 test event classes from TRECVID 2011 MED. The details of this experimental results, in addition to the comparison of performance with and without the use of non-target scores illustrated in Fig. 2 are described in Sec. 5.

5 Experiments and Result Analysis

We have applied the proposed framework on two challenging large-scale consumer video datasets including TRECVID '11 MED [3] and Columbia Consumer Video (CCV) [2] datasets. Both the size and complexity of the datasets are beyond other alternatives such as YouTube Sports [11] or Holywood datasets [13].

Our proposed methods are compared against other standard fusion techniques [10,4] based on logistic regression (LR) and linear SVM. We also compare fusion results with and without non-target base classifier scores, as discussed in Sec. 4.2. For all experiments, performance measure in Eq. 1 has been used, with $\tau = 12.5$ for TRECVID '11 MED and $\tau = 10$ for the CCV dataset. For the training of comparative approaches, we have assigned the weights equal to τ to positive samples. Operating points were selected on the training performance curves where the specified ratio τ is satisfied. Finally, the performance metrics are computed at the selected operating points.

5.1 Results on TRECVID 2011 MED dataset

TRECVID 2011 MED corpus [3] provides an excellent test-bed for real-world video retrieval problems due to its large size (45K video clips) and huge inter- and intra-class content variability. For the MED task, there are 10 annotated

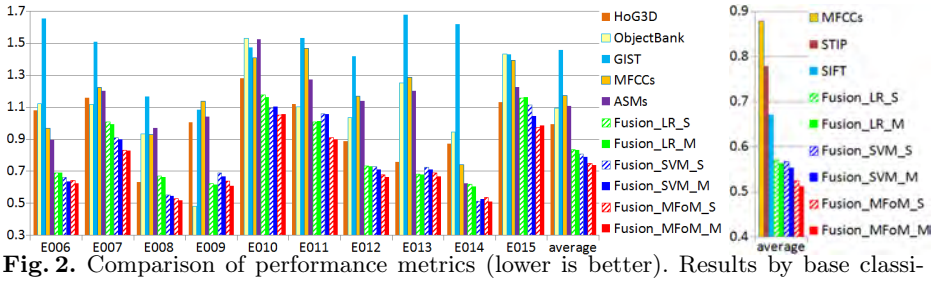

Fig. 2. Comparison of performance metrics (lower is better). Results by base classifiers, LR-, SVM-, and MFoM-based fusion with only target class scores ('_S') and additional non-target class scores ('_M') are shown; **(Left)** 10 classes and average from TRECVID 2011 MED and **(Right)** Average of 20 classes from the CCV dataset

event classes: E006-*Birthday party*, E007-*Changing a vehicle tire*, E008-*Flashmob gathering*, E009-*Getting a vehicle unstuck*, E010-*Grooming an animal*, E011-*Making a sandwich*, E012-*Parade*, E013-*Parkour*, E014-*Repairing an appliance*, and E015-*Working on a sewing project*.

Features and Base Classifiers. We used five types of features in our experiments: HoG3D [14], Object Bank (OB) [15], GIST [16], MFCCs [17], and ASMs [17]. HoG3D is designed to be a low-level feature which captures texture and motion within videos, while OB consists of 177 semantic object detectors. MFCCs and ASMs are low- or mid-level audio features. For HoG3D, MFCC, and ASMs, we aggregated them into a clip-level bag-of-words feature, and learned a SVM with a histogram intersection kernel (HIK). For OB, we aggregated the features using max-pooling across multiple frames, and learned a linear SVM. For GIST, we learned a linear SVM using per-frame features, and performed clip-level classification by averaging the scores over multiple frames (Note this paper is not focused on the specifics of the base classifiers, but rather on their fusion).

Comparison of fusion performance on the target metric. The overall performance is summarized in Fig. 2(Left) where lower bars indicate superior performance. For training of different fusion classifiers (MFoM, SVM, LR), identical base classifier scores were used where the results with and without non-target class scores are denoted by postfixes `_M` and `_S` respectively. It can be observed that Fusion_MFoM_M (S_τ=0.7374) achieves the best performance consistently across all events, where it shows meaningful improvement of relatively 12.9% on average, against Fusion_LR_M (S_τ=0.8326) and 7.3% from Fusion_SVM_M (S_τ=0.7916). A similar result holds when using only target-class scores (`_S`).

The benefits of explicit performance metric optimization by our methods can be examined in more detail by looking at the the detection error tradeoff (DET) curves [18] for three test event classes shown in Fig. 3(Left). For the three event classes, the DET curves of the proposed MFoM approach (red) is superior or comparable to the other approaches. The remaining seven event classes showed similar patterns. However, while MFoM performs better than the other methods around the operating point, it is not always better away from the operating point (e.g. E15 (solid) and E07 (dot-dash)). This is not unexpected, since the goal of our approach is to explicitly improve performance at the operating point.

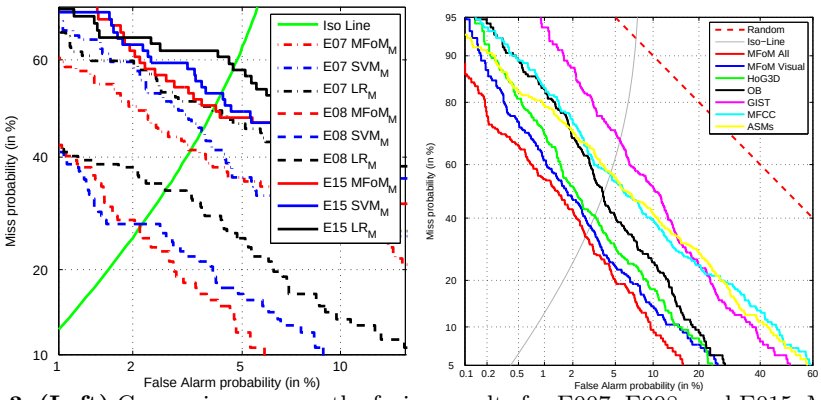

Fig. 3. (Left) Comparison among the fusion results for E007, E008, and E015. MFoM outperforms SVM and LR, especially along with the isoline of $P_{\mathrm{FA}} : P_{\mathrm{MD}} = 1 : 12.5$. **(Right)** Comparison among the results by the base and fusion classifiers for E012.

In terms of training parameters, the MFoM fusion classifiers were trained with the following parameters: $\alpha = 30$, and $\gamma = 0.2 \sim 0.4$. γ varies across classes, and was determined by cross-validation. Similar cross validation schemes were used to identify optimal parameters for SVM and LR.

Performance of base classifiers and the effect of fusion. Among the individual features shown in Fig. 2(Left), HoG3D shows the best performance on average, especially for events with temporal dynamics, such as E012. Next, OB is followed, which is competitive for relatively static classes, such as E011. Notably, audio features are competitive for audio-rich events such as E006.

All the fusion methods consistently outperform base classifiers, showing the clear benefits of fusion. For example, Fig. 3(Right) illustrates the effect of fusion by the proposed algorithm for E012 in the DET plot. It is notable that the fusion of the visual features (blue line) is better than the individual visual features (HoG3D, OB, and GIST). Furthermore, the final fusion result (red line) is improved by additionally incorporating the audio features.

For a qualitative assessment, Fig. 4 shows the top retrieved results for E011 from the proposed fusion approach and the two base classifiers. It is interesting to see that the two visual features seem complementary. HoG3D captures textures of scenes as well as temporal dynamics, while OB outputs responses from object detectors. Accordingly, some of the top results by HoG3D are mainly triggered by only textures such as roads or plain background, while most of the top results by OB contain a vehicle in the middle. Combining textures of scenes and responses from object detectors, the fusion results show much better performance that mostly have a vehicle object and a consistency in spatio-temporal dynamics.

Model parameters and effect of additional non-target class scores. The learned model parameters of our MFoM fusion scheme are illustrated in Fig. 5. Each row represents 50-dimensional LDF parameters, which is composed of the weights for the 10-dimensional scores from each feature block. A high positive value indicates strong positive correlation of the corresponding score element to a target class, while a negative value implies a negative correlation. Diagonal structures are observed because base classifiers learned for the same target class

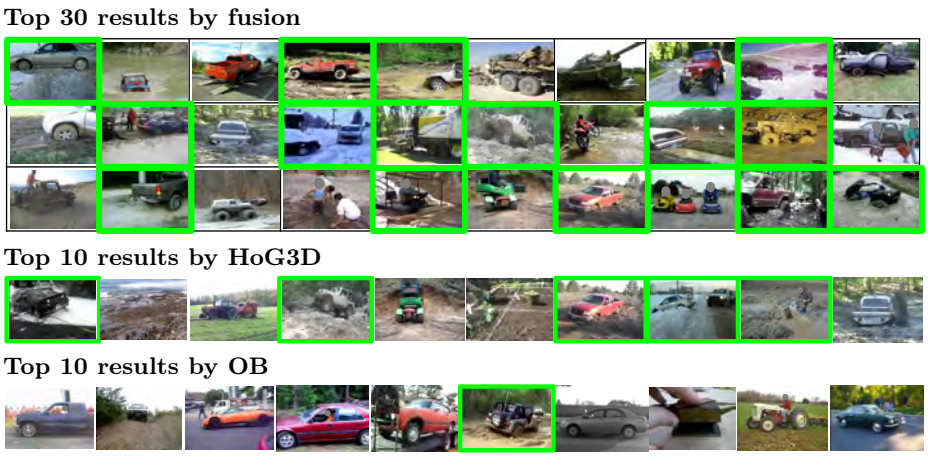

Fig. 4. Top 30 results by the proposed fusion algorithm, top 10 results by HoG3D and OB; sorted from top-left to botom-right. True positives are marked with green boxes.

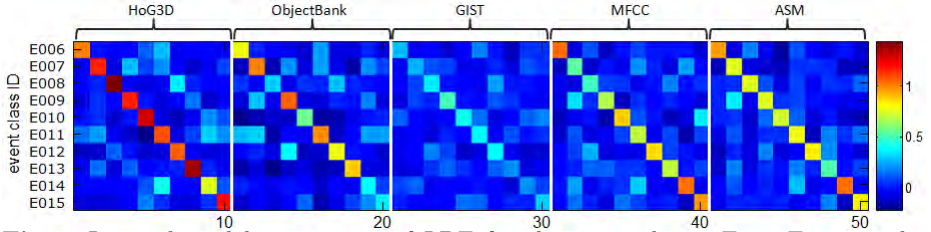

Fig. 5. Learned model parameters of LDF for the event classes E006–E015 on the MED dataset. Each row is the 50-dimensional model parameter of one-versus-all fusion classifiers for every event. Each column corresponds to one of 50 base classifiers.

are more discriminative, as expected. It is also interesting to see correlations between different event types. For example, the fusion classifier for E011 (row 6) shows positive correlation with ObjectBank base classifiers (column 11) for E006, perhaps because both events frequently occur in dining rooms.

5.2 Results on Columbia Consumer Video dataset

As the second dataset, we applied the proposed fusion scheme on Columbia Consumer Video (CCV) dataset [2], which is another publicly available large-scale consumer video dataset. It includes 9,317 consumer videos in 20 complex event classes. In addition, it provides 3 types of precomputed bag-of-words features for SIFT, STIP [13], and MFCC.

We conducted identical experiments on CCV dataset. For all three types of features, base classifiers are learned using HIK SVMs. Then, LR-, SVM-, MFoM-based fusion classifiers were learned on top of the identical base classifiers. Experimental results on CCV dataset are summarized on Fig. 2(Right). Patterns identical to the results on TRECVID dataset has been observed for all 20 event classes. For brevity, only the average performance across all classes is shown here. Overall, there is an average gain of 10.1% and 6.3% achieved by MFoM fusion (MFoM_M, S_τ=0.5208), over the LR fusion method (LR_M, S_τ=0.5637) and the SVM fusion method (SVM_M, S_τ=0.5536), respectively.

6 Conclusion

In this work, we have presented our novel late-fusion framework for video retrieval, which explicitly optimizes given performance metrics. In particular, we showcased an effective approximation scheme for the important class of weighted metrics which can include sub-metrics such as P_{MD} and P_{FA}, and requirements for an operating point. Our experimental results on two large consumer video archives are promising, and suggest that our approach will add value for real-world computer vision applications with sophisticated user needs.

References

1. Toderici, G., Aradhye, H., Pasca, M., Sbaiz, L., Yagnik, J.: Finding meaning on youtube: Tag recommendation and category discovery. In: CVPR. (2010)
2. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: ACM ICMR. (2011)
3. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: ACM MIR. (2006)
4. Wang, Z., Zhao, M., Song, Y., Kumar, S., Li, B.: Youtubecat: Learning to categorize wild web videos. In: CVPR. (2010)
5. Yang, W., Toderici, G.: Discriminative tag learning on youtube videos with latent sub-tags. In: CVPR. (2011)
6. Joachims, T.: A support vector method for multivariate performance measures. In: ICML. (2005)
7. Calonder, M., Lepetit, V., Fua, P.: Pareto-optimal Dictionaries for Signatures. In: CVPR. (2010)
8. Gao, S., Wu, W., Lee, C.H., Chua, T.S.: A mfom learning approach to robust multiclass multi-label text categorization. In: ICML. (2004)
9. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: ICCV. (2007)
10. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: IEEE International Conference on Computer Vision (ICCV). (2009)
11. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos ïn the wild:. In: CVPR. (2009)
12. Katagiri, S., Juang, B.H., Lee, C.H.: Pattern recognition using a family of design algorithm based upon the generalized probabilistic descent method. Proc. of the IEEE (1998) 2345–2373
13. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
14. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008)
15. Li, L.J., Su, H., Xing, E.P., Fei-Fei, L.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: Proceedings of the Neural Information Processing Systems (NIPS). (2010)
16. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. **42** (2001) 145–175
17. Lee, C.H., Soong, F., juan, B.H.: A segment model based approach to speech recognition. In: ICASSP. (1988)
18. Martin, A.F., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of detection task performance. In: Eurospeech. (1997)