PER-EXEMPLAR FUSION LEARNING FOR VIDEO RETRIEVAL AND RECOUNTING

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

[†]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[‡]Kitware Inc., Clifton Park, NY, USA

ABSTRACT

We propose a novel video retrieval framework based on an extension of per-exemplar learning [7]. Each training sample with multiple types of features (e.g., audio and visual) is regarded as an exemplar. For each exemplar, a localized per-exemplar distance function is learned and used to measure the similarity between itself and new test samples. Exemplars associate only with sufficiently similar test data, which accumulate to identify the data to be retrieved. In particular, for every exemplar, relevance of each feature type is discriminatively analyzed and the effect of less informative features is minimized during the fusion-based associations. In addition, we show that our framework can enable a rich set of recounting capabilities where the rationale for each retrieval result can be automatically described to users to aid their interaction with the system. We show that our system provides competitive retrieval accuracy against strong baseline methods, while adding the benefits of recounting.

Index Terms— video retrieval, video recounting, fusion

1. INTRODUCTION

Most semantic categories used for multimedia retrieval have inherent within-class diversity. For example, consider a video search for the concept class 'wedding ceremony'. Across different cultures, both their looks (visual) and music (audio) are fairly different. The diversity can be dramatic, and it raises the question whether the conventional approaches (e.g., [3, 11]) which learn a single classifier per category can still be successful and scalable as the number of concepts and diversity increase. Furthermore, when blackbox methods such as SVMs are used, recounting of search results or in-depth analysis of the underlying training data has been challenging. By recounting, we mean the ability to automatically explain to the users why some results are retrieved at all, and what particular characteristics triggered them to be returned as results, which is a core high-level challenge the multimedia community needs to address.

In this study, we propose a retrieval framework based on per-exemplar fusion associations, as a solution to address the aforementioned challenges. Our approach regards training samples as exemplars, and learns localized per-exemplar distance functions centered around each sample. In this way, all the diversity within training data is maintained in a straightforward manner. For a new sample, each local distance function only associates itself with samples that are sufficiently similar. Thus, the notion of retrieval is re-defined as an association problem where test data with relatively high ratio of positive associations are retrieved.

In particular, our framework is designed to incorporate and fuse multiple types of heterogeneous features, with an emphasis on video retrieval problems. Overall, the resulting learning architecture can be understood as a non-parametric variant of late-fusion approaches where discriminative per-feature base classifiers are used. In detail, an association between two samples is established by a set of distances across different feature types. Furthermore, for every training exemplar, the relevance of each feature is measured based on its discriminative power around its neighborhood. This is particularly useful because some features may be more relevant for certain exemplars. For example, imagine a birthday video clip recorded in a dark room with a crowd singing a birthday song. It is crucial to learn that the audio (and not visual appearance) is the main relevant feature and similar samples are discovered mostly based on audio. We show that the per-exemplar relevance of each feature as well as the importance of each exemplar can be automatically analyzed and incorporated to achieve competitive retrieval accuracy.

In addition, we show that our framework enables a rich set of recounting or summarization capabilities. Due to the nature of association-based retrieval, it is straightforward to identify the exemplars which actually triggered on the retrieved data. In addition, existing knowledge related to relevant features can be transferred from the exemplars to the target data to describe it. For example, if a large number of exemplars with a metadata tag (e.g., dynamic motion or rock music) associate with a clip, the metadata can be used to au-

We thank Anthony Hoogs, Greg Mori, Arash Vahdat, Zhi Feng Huang, Weilong Yang, Scott McCloskey, Ben Miller, and Byungki Byun for helpful comments and sharing features used in this work. This work is 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 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.

tomatically describe the new data. Furthermore, the relevance of each feature dimension can be used to indicate the core evidence considered during the retrieval process. For example, users can understand easily that a particular result has been retrieved due to its audio and/or visual evidence.

Our work builds upon past work on association-based visual object recognition for images [1, 7, 8] and discriminative nearest neighbors methods [2, 12]. Compared to [1, 7, 8], our method identifies relevance of features for exemplars based on discriminative analysis. Furthemore, our work, with extensive use of both audio and video features for unconstrained consumer video retrieval, provides a new perspective beyond relatively constrained image retrieval problems. Compared to [2, 12], our approach yields explicit and well-defined normalized distance functions and provides a more principled manner to compute classification probabilities for new test samples, along with additional recounting capabilities.

The empirical usefulness of the proposed method is evaluated on a challenging real-world dataset where we demonstrate competitive retrieval accuracy which matches or exceeds other conventional approaches. Furthermore, aforementioned novel recounting aspects of our method is highlighted through qualitative analysis.

2. PER-EXEMPLAR SIMILARITY

In this work, each training sample is regarded as an exemplar, and a local distance function is defined for each exemplar to measure similarities of samples to the exemplar. Let $F = \{f_i \,|1 \leq i \leq N\}$ be a set of N types of available features, where each feature is a multi-dimensional vector.

2.1. Local distance function

The local distance $D_e(s)$ from an exemplar e to a test sample s can be measured by aggregating a set of feature-wise elementary distances $d_i(e,s)$ computed for each f_i. We use a linear combination, as follows:

$$D_e(s) = \sum_{i=1}^{N} \omega_i(e) \times d_i(e,s) = \langle \omega(e) \cdot d(e,s) \rangle, \quad (1)$$

where $\omega_i(e) \geq 0$ denotes the relevance weight for the i-th feature and the corresponding elementary distance $d_i(e,s)$.

Accordingly, each per-exemplar distance function is characterized by a $1 \times N$ parameter vector $\omega(e)$. With a higher value of the weight $\omega_i(e)$, the local distance function is more heavily influenced by the similarity in the i-th feature, and vice versa. The non-negative condition for weights is imposed to ensure that larger elementary distances always lead to larger overall aggregate distances and maintains the notion of distance, which is advocated also in [1, 7].

2.2. Discriminative elementary distance

Among many possible choices, in our per-exemplar fusion approach, we employed discriminative scores estimated by base classifiers to compute an elementary distance for each feature. First, for each feature type f_i, we learn a discriminative base classifier per concept in a one-vs-all manner. Then, we derive an elementary distance $d_i(e,s)$ from the scores output by the discriminant function $g_i(\cdot)$ learned from the i-th feature. ($g_i(\cdot)$ measures the confidence of its input matching a target class based on feature type f_i.)

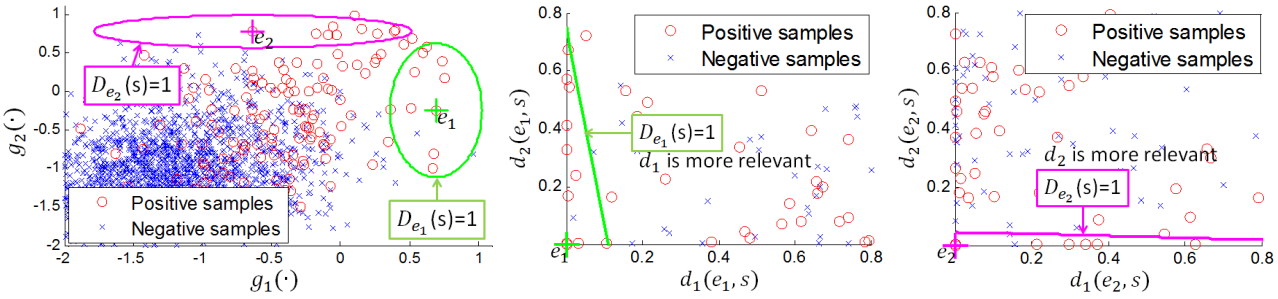
Specifically, we use the squared difference between discriminative scores:

$$d_i(e,s) = |g_i(e) - g_i(s)|^2 . \quad (2)$$

Hence, with Eq. 2, a local distance between an exemplar e and a sample s is measured, which is mapped from a feature space to a discriminative score space by the base classifiers. With Eqs. 1 and 2, a local distance function, which is a linear combination of elementary distances, appears as an ellipsoid centered around the corresponding exemplar in a discriminative score space. We observed that our approach using the discriminative elementary distance can improve the retrieval accuracy, showing results superior to other approaches using generative elementary distance such as L_2 distance as studied in [1, 7], especially when a feature vector is high dimensional and extremely sparse, e.g., bag-of-words (BoW) features with thousands codewords.

By using the elementary distance based on discriminative scores, we could acquire more discrimination power than using generative distance such as L_2 distance, We can also take advantage of kernelization in training base classifiers, which often shows significant improvements in many multimedia applications. Moreover, the compact representation by discriminative scores can provide a robust approach to fuse multiple types of heterogeneous features by transforming them into a common vector space.

Fig. 1 geometrically illustrates the local distance functions $D_{e_1}(s)$ and $D_{e_2}(s)$ of the two exemplars e_1 (green) and e_2 (magenta). For clarity, only two types of features are considered in this example. In Fig. 1(a), positive (red 'o') and negative (blue 'x') samples are scattered in a two dimensional discriminative score space by their confidence measures from discriminant functions $g_1(\cdot)$ and $g_2(\cdot)$ learned from different features. Figs. 1(b) and (c) are drawn in the elementary distance space defined in Eq. 2 from the two exemplars e_1 and e_2, respectively. (Each exemplar is located at the origin since $d_i(e,e) = 0$.) In this example, we gave the exemplar e_1 a higher relevance weight for the elementary distance $d_1(\cdot)$ than $d_2(\cdot)$, while the exemplar e_2 had a higher relevance weight for the elementary distance $d_2(\cdot)$ than $d_1(\cdot)$. It is clear that the iso-local distance line $D_{e_1}(s) = 1$ of the exemplar e_1 in Fig. 1(b) is steeper compared to $D_{e_2}(s) = 1$ of e_2 in Fig. 1(c). In addition, the iso-local distance lines in the elementary distance spaces appear as the ellipses with the different ratios in Fig. 1(a). A higher relevance weight makes the local distance more heavily influenced by the similarity in the corresponding feature; for example, the iso-local distance ellipse of e_2 contains samples which share more similar scores from the second feature ($g_2(\cdot)$-axis) than the first feature ($g_1(\cdot)$-axis).



(a) Distribution of discriminative scores (b) Elementary distance from e_1 (c) Elementary distance from e_2

Fig. 1. The distribution of discriminative scores (a), and the elementary distance spaces centered around two different exemplars (b) and (c), respectively. Each exemplar has its own local distance function with different relevance weights, where the iso-local distance lines with the different slopes in (b) and (c) appear as the iso-local distance ellipses with the different ratios in (a).

In this manner, we can determine which feature is more relevant to the similarity with a given exemplar than other features. This is one of the key aspects of the proposed per-exemplar fusion algorithm, and examples for video recounting will be discussed in Section 4.

3. RETRIEVAL BY LOCAL DISTANCE FUNCTION

In the following sections, we present our approach towards learning relevance weights in Eq. 1 for features of each exemplar. At a high-level, relevance weights of an exemplar e are learned to assign small distance to neighboring samples with the same class with e, while assigning large distance to all the competing samples. In addition, we show how the learned local distance functions can be combined to estimate the classification probabilities for test samples.

3.1. Learning feature relevance weights

In our approach, the set of feature relevance weights for an exemplar e belonging to a particular concept class C is learned by an iterative discriminative neighborhood analysis. During learning, we incorporate the set of the most similar K positive examples $S_e(C,K)$ and all the available negative examples $\overline{S(C)}$. Accordingly, training samples for each exemplar is different, and can be denoted as $S_e = S_e(C,K) \cup \overline{S(C)}$. The use of only the nearest positive samples is to ensure that localized relevance can be learned, differently per exemplar. Here, the set of positive nearest samples $S_e(C,K)$ are found based on the distance function $D_e(\cdot)$ whose parameters $\omega(e)$ are iteratively updated to maximize discrimination among training samples. Accordingly, the learning process can be understood as the simultaneous estimation of the weight vector $\omega(e)$ and the localized training subset $S_e(C,K)$. We formulate this overall iterative optimization problem as the following max-margin learning problem with hinge loss functions:

$$\{\omega'^*(e), S_e^*\} = \text{argmin}_{\omega'(e),S_e} f(\omega'(e), S_e) \quad (3)$$

$$f(\omega'(e), S_e) = \tfrac{1}{2}\|\omega'(e)\|^2 + c_1 \textstyle\sum_{j\in S_e(C,K)} \xi_j + c_2 \sum_{j\in\overline{S(C)}} \xi_j$$
$$s.t.\ \forall i,j : \langle \omega'(e) \cdot d'(e,s_j)\rangle \geq 1-\xi_j, \xi_j \geq 0, \omega_i(e) \geq 0, \quad (4)$$

where the two extended vectors $\omega'(e) = [\beta_e; \omega(e)]$ and $d'(\cdot) = [1; d(\cdot)]$ are used to consider a bias term, and constant parameters c_1 and c_2 control the effect of loss terms from the K most similar and competing samples, respectively.

Given a training subset S_e, minimizing Eq. 4 is a conventional convex programming problem which can be solved by a quadratic programming (QP) method such as SVM with the additional non-negative constraints, $\forall i : \omega_i(e) \geq 0$. Then, we can solve Eq. 3 iteratively. Starting from an initial relevance weight vector $\omega'^0(e)$, the current training subset S_e^k can be found based on a previous $\omega'^{k-1}(e)$, from which we estimate the current parameters $\omega'^k(e)$ by minimizing the cost function $f(\omega'^k(e), S_e^k)$ at every k-th iteration until the K most similar set $S_e(C,K)$ converges.

Solving Eq. 3 can be considered as a process of finding a decision boundary between similar neighbors with the same class as an exemplar e and all competing samples with different class labels. After acquiring the optimal relevance weights $\omega(e)$, we divide it by the bias term β_e to normalize so that the value of a learned local distance function at the decision boundary becomes 1. Then, samples within the boundary are defined as 'associated' samples to the exemplar e. Due to the normalization, distances to all associated samples are $[0,1)$. Since a decision boundary learned in Eq. 4 is characterized by neighboring training samples of each exemplar, the number of associated test samples may also differ for each exemplar, while a conventional k-NN method keeps the same number of associations across all samples.

3.2. Probability estimation for retrieval

In this work, we design the probability $\hat{P}(C|s)$ of a test sample s being in a class C to be estimated based on the general form shown in Eq. 5, where the set of all exemplars that built associations with the test sample s is denoted by A_s, while a subset of them which are positive samples of class C is denoted by A_s^C.

$$\hat{P}(C|s) = \frac{\epsilon + \sum_{e\in A_s^C} h(D_e(s))}{\epsilon + \sum_{e\in A_s} h(D_e(s))} \quad (5)$$

Above, the influence function $h(\cdot)$ represents a class of arbitrary functions that decrease w.r.t. the distance between an

exemplar e and the a target sample s. Intuitively, Eq. 5 states that the probability is estimated as the sum of influence by the positive examples, divided by the total amount of influence by all associations. The term ϵ provides smoothing prior which is helpful when the number of associations are sparse or heavily skewed. We found that the proposed distance-based probability model can improve the retrieval accuracy for reasonable choices for h, delivering results superior to the conventional approaches using $\hat{P}(C|s) = |A_s^C|/|A_s|$ which is employed by methods such as k-NN.

4. EXPERIMENTAL RESULTS

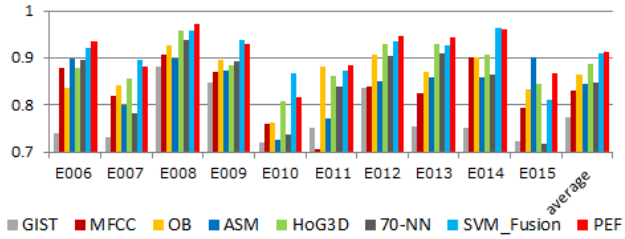
To assess the proposed method, we conducted experiments on a challenging real-world video dataset. For our experiments, we employed the TRECVID 2011 multimedia event detection (MED) data [10]. The MED data provides an excellent test-bed for real-world video retrieval and recounting problems due to its large size and diversity. It consists of consumer videos on the Internet. Accordingly, huge within-class content variability poses significant challenges and fusion can improve retrieval. It consists of 13K training and 32K test video clips labeled with 10 event classes and a pure negative class. The 10 event classes are follows: *Birthday party* (E006), *Changing a vehicle tire* (E007), *Flash mob gathering* (E008), *Getting a vehicle unstuck* (E009), *Grooming an animal* (E010), *Making a sandwich* (E011), *Parade* (E012), *Parkour* (E013), *Repairing an appliance* (E014), and *Working on a sewing project* (E015). For each event class, the training data is substantially imbalanced: there are ~150 positive training samples on average per class, while there are more than 11K pure negative training video clips in total. In the test data, which consists of 32K clips, there are ~120 positive examples (0.4 percent) for each class on average, and approximately 31K videos were pure negative.

4.1. Features and Discriminative distances

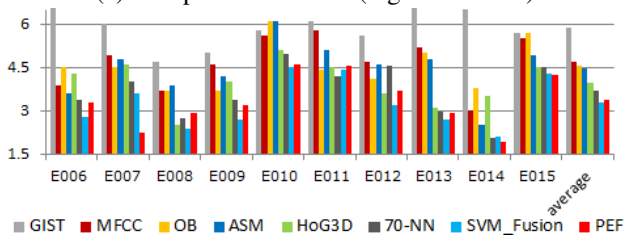
To capture diverse information from videos, total of 5 different features were computed. They include 3 visual and 2 audio features: HoG3D bag-of-words (BoW) [4], Object Bank (OB) [6], GIST [9], MFCC BoW, and acoustic segment models (ASMs) BoW [5]. For each feature, standard SVMs are individually learned as base classifiers in a one-vs-all manner for each event class. Then, they are used on test data to generate scores which are used as basis to measure per-feature discriminative distance between samples.

HoG3D BoW+HIK SVM: HoG3D [4] is a spatio-temporal extension of histogram of gradients (HoG), which captures additional motion information. Once features are extracted regularly across frames, they are quantized to 1,000 codewords and used to form HoG3D BoW features. SVMs with histogram intersection kernels (HIK) were used.

OB+Linear SVM: OB features are designed to capture semantic visual signals by a set of object detectors. Total of 171 object detectors provided by an existing implementation



(a) Comparison of AUC (higher is better)



(b) Comparison of $P_{FA}(\%)$ at TER (lower is better)

Fig. 2. Performance comparison for video retrieval by different approaches including base classifiers without fusion, k-NN, fusion by a single SVM, and per-exemplar fusion. Two metrics are used, including (a) area under curve (AUC), and (b) P_{FA} at TER.

[6] were applied on frames, with sliding windows at multiple scales. Clip-level features are computed using max-pooling across frames. Then, linear SVMs were trained.

GIST+Linear SVM+Average: GIST features [9] capture contextual scene appearance. It can provide discrimination among different types of scenes such as outdoor versus indoor. A linear SVM was trained on per-frame features, while the scores in the test phase are computed across frames and averaged to provide a single score for a clip.

MFCC BoW+HIK SVM: MFCCs were incorporated as low-level audio features to capture patterns of sound embedded in videos. HIK SVMs have been used on a clip-level BoW feature with a 1024-sized codebook.

ASM BoW+HIK SVM: ASMs [5] were also incorporated to capture more semantic acoustic information. ASMs are composed of audio detectors for characteristic sound, e.g., music, laughing, clapping, phonetic sounds, and etc. A clip-level audio is decoded with 165 ASMs learned by 16 types of sounds and 39 phonemes. Then, ASM BoW was formulated with unigrams and bigrams. HIK SVMs were used.

4.2. Video Retrieval Performance and Comparisons

We conducted evaluation of the proposed per-exemplar fusion (PEF) algorithm for video retrieval, and compared it with other methods including per-feature base SVM classifiers, and two alternative fusion methods: (1) k-nearest neighbors (k-NN) and (2) a standard SVM fusion. For k-NN, a score on a test clip was computed by the proportion of the positive neighbors among k neighbors, where we used $k = 70$ (found by cross validation). NNs are found by unweighted Euclidean distances based on discriminative distances, which is a special case of our approach. For SVM fusion, a single SVM fusion classifier was trained using score features formed

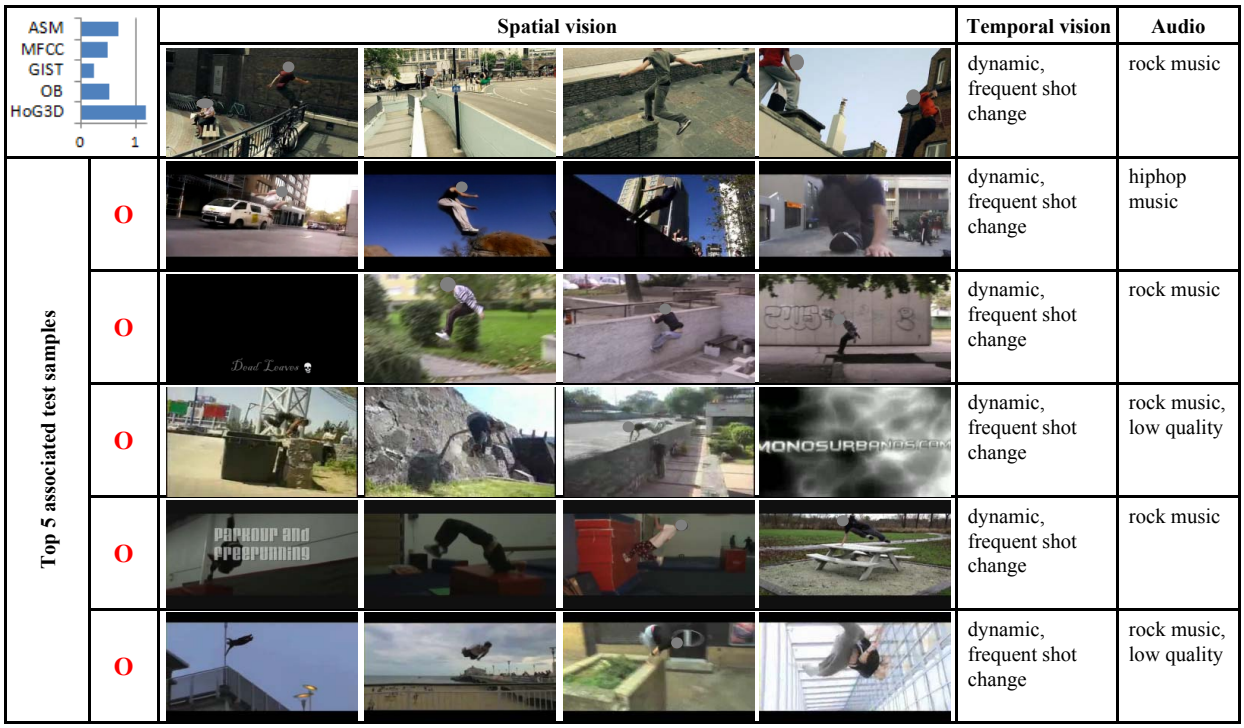


Fig. 3. Example of video recounting for *Parkour*: HoG3D is most significant for associations with the given training exemplar, which contains fast movements of objects and frequent shot changes. The consistent music sound type is also notable.

by concatenating all available discriminative base classifier scores. Other non-discriminative distances have been studied, but, their results were inferior. Accordingly, we have not pursued the direction further, which is omitted for brevity.

Two performance measures were adopted for our evaluation: (1) area under ROC curves (AUC), and (2) the probability of false alarms (P_{FA}) at a target error ratio (TER) of P_{FA} over the probability of missed detections (P_{MD}). A specific TER can capture an aspect of user experience regarding the tolerance they are willing to assume between a missed detection and a false alarm. The TER was set to be $P_{FA} : P_{MD} = 1 : 12.5$ in this work. Note that, because positive samples constitute only 0.4 percent of the test data, the per-sample mis-classification cost is still ~16 times higher for a positive sample than a negative one. In terms of implementation details, the following parameters were used for PEF, based on the analysis of data label imbalance and cross-validation: $\{K, \alpha, \epsilon, c_1, c_2\} = \{5, 1.5, 20, 1\}$. For the influence function in Eq. 5, the following variant of Gaussian function has been used for the result reported in this work: $h\left(D_e\left(s\right)\right) = \exp\left[-\alpha\left\{D_e\left(s\right)\right\}^2\right]$.

A summary of the classification results for the 10 test classes are shown in Fig. 2. For the two metrics, it can be observed that PEF and SVM fusion consistently outperform all base classifiers, while k-NN shows degradation in AUC. Between PEF and SVM fusion, we observe comparable classification performances: PEF (0.9138) is slightly better than SVM fusion (0.9089) in AUC, but SVM fusion (3.27%) is slightly better than PEF (3.36%) in P_{FA} at TER. Given additional advantages provided by PEF, such as recounting capabilities, the top-end video retrieval performance is appealing.

4.3. Qualitative Analysis and Recounting

In addition to favorable retrieval performance, PEF provides notable advantages regarding recounting, which is enabled by the association-based retrieval scheme. We can look into the learned relevance weights of per-exemplar local distance functions and obtain insights about the core characteristics of the exemplar and their associations. In general, samples are associated when highly weighted features are similar to the exemplar. Two examples are shown in Fig. 3 and 4 where the learned relevance weights are visualized for the training examples at the top, along with the top-ranked associated test examples below. The test examples that share identical labels with the exemplar are marked by circles, otherwise, by crosses. Additionally, frames from each video clip are shown, along with manually marked visual and audio characteristics.

Fig. 3 illustrates an example video for *Parkour*. The top row indicates that the given training exemplar will associate with samples that have high correlation for HoG3D followed by ASM. The top 5 associated test samples indeed show significant temporal dynamics, i.e., jumping, running fast, dumbling, and etc. It can be observed that the potential automatic recounting of those examples by transferring both temporal vision and audio characteristics of the exemplar will be fairly accurate, regardless of the diversity in the data.

Another example for *Grooming an animal* is illustrated

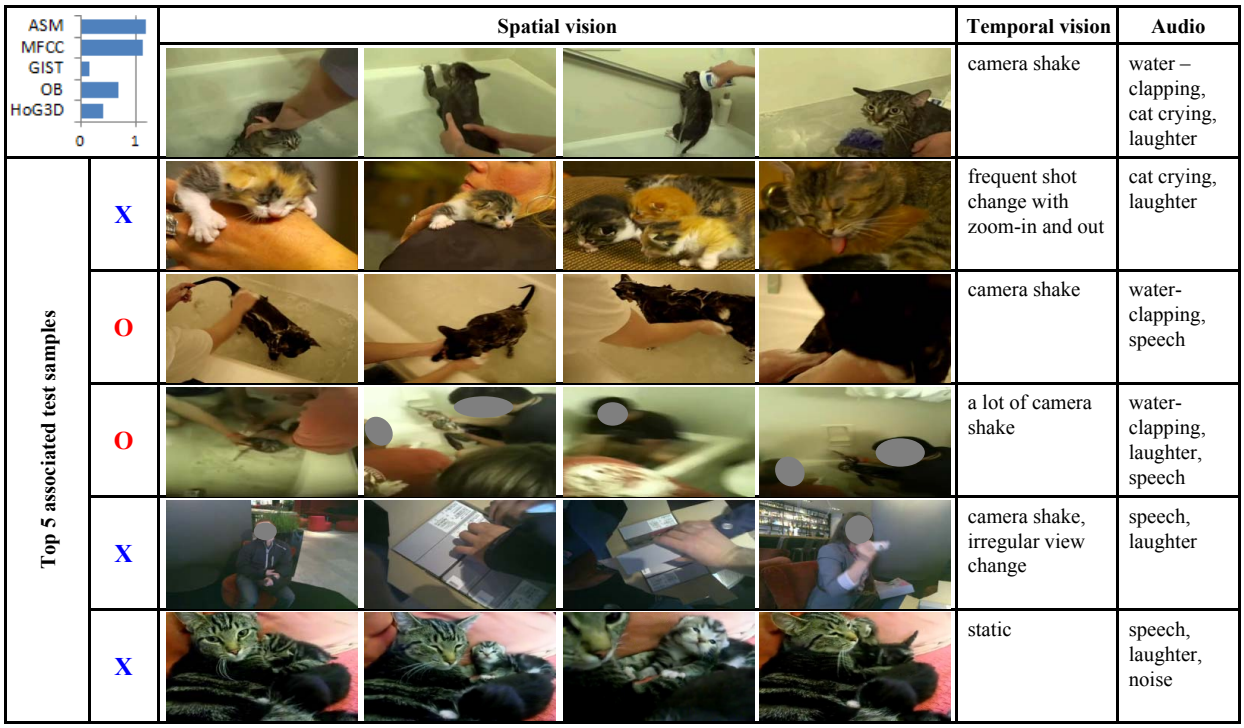


Fig. 4. Example of video recounting for *Grooming an animal*: Both audio features are significant, triggered by water-clapping sound for associations with the given training exemplar, which includes water-clapping and laughter sound types.

in Fig. 4. In this exemplar, audio evidence such as water-clapping, cat crying, speech, and laughter are strong, and it can be observed that the corresponding audio features are highly weighted. The 3rd associated sample contains very blurry spatial vision due to camera shakes; however, it could be still accurately retrieved based on audio evidence. While the accuracy of associations is limited in terms of the class label, it is notable that even the incorrect results share similar audio properties. Among visual features, OB is the most important, and it can be seen that associated examples indeed contain similar objects such as cats and hands.

5. CONCLUSION

We presented our novel PEF framework for video retrieval and recounting. Our approach incorporates novel schemes to learn and use discriminative local distance functions to associate with examples that share similar properties on core feature channels. Our experimental results on a large consumer video archive is promising: (1) PEF shows favorable retrieval results comparable to competetive alternatives, (2) Furthermore, PEF provides substantial advantages towards understading core characteristics of each exemplar and automatic tagging of associated samples and retrieval results, which straightforwardly leads to detailed recounting.

6. REFERENCES

[1] A. Frome and Y. Singer. Image retrieval and classification using local distance functions. In *NIPS*, 2006.

[2] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *PAMI*, 18(6):607–616, 1996.

[3] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM ICMR*, 2011.

[4] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[5] C.-H. Lee, F. Soong, and B.-H. Juang. A segment model based approach to speech recognition. In *ICASSP*, 1988.

[6] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[7] T. Malisiewicz and A. A. Effros. Recognition by association via learning per-exemplar distances. In *CVPR*, 2008.

[8] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *ICCV*, 2011.

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[10] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *ACM MIR*, 2006.

[11] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal Multimodal Fusion for Multimedia Data Analysis. In *ACM Multimedia*, 2004.

[12] H. Zhang, A. C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In *CVPR*, 2006.