Segmental Multi-way Local Pooling for Video Recognition

Ilseo Kim
Kitware Inc.
ilseo.kim@kitware.com

Sangmin Oh
Kitware Inc.
sangmin.oh@kitware.com

Arash Vahdat
Simon Fraser University
avahdat@sfu.ca

Kevin Cannons
Simon Fraser University
kcannons@sfu.ca

A.G.Amitha Perera
Kitware Inc.
amitha.perera@kitware.com

Greg Mori
Simon Fraser University
mori@cs.sfu.ca

ABSTRACT

In this work, we address the problem of complex event detection on unconstrained videos. We introduce a novel multi-way feature pooling approach which leverages segment-level information. The approach is simple and widely applicable to diverse audio-visual features. Our approach uses a set of clusters discovered via unsupervised clustering of segment-level features. Depending on feature characteristics, not only scene-based clusters but also motion/audio-based clusters can be incorporated. Then, every video is represented with multiple descriptors, where each descriptor is designed to relate to one of the pre-built clusters. For classification, intersection kernel SVMs are used where the kernel is obtained by combining multiple kernels computed from corresponding per-cluster descriptor pairs. Evaluation on TRECVID '11 MED dataset shows a significant improvement by the proposed approach beyond the state-of-the-art.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; Information Search and Retrieval.

Keywords

Representation, Multimedia Event Recognition, TRECVID

1. INTRODUCTION

Detection of complex events on unconstrained real-world videos (e.g., YouTube) is a challenging problem. Most complex events (e.g., *birthday party* and *board trick)* exhibit large intra-class variations, and videos frequently consist of multiple segments exhibiting different and evolving contents that include not only a mixture of contents closely related to events but also temporal clutters such as title screens or irrelevant contents arbitrarily stitched-in by users.

Many reported successful retrieval systems (e.g., [3, 6]) for unconstrained videos share the common idea of constructing clip-level representations via global average-based pooling. For each feature type, a bag-of-words (BoW) descriptor (or variations) is built per video by pooling across the entire video. These globally pooled features work well, although the fact that these methods do not exploit detailed segment information leaves a room for further research, which started to be addressed by recent efforts such as [2, 10].

In this work[1], we present a multi-way local pooling (MLP) approach that leverages the detailed segment-level information and boosts the performance beyond the globally pooled descriptors. The overall scheme is illustrated in Fig.1. Our approach builds multiple descriptors per video, where each descriptor is designed to relate to one of the pre-built segment clusters. These clusters are constructed in an unsupervised manner and can be understood as rough themes interchangeably appearing as segments in videos. From an input video, a separate descriptor is built per cluster by accumulating features from segments which are *local* (or similar) to the represented cluster. The rationale behind the MLP strategy is partly inspired by the recently introduced theory of local pooling [1] which showed that pooling features similar in multi-dimensional input space separately improves the representational power and classification accuracy. In addition, we observe that the frequency of segment-to-cluster assignments provides a unique signature to indicate the importance of each cluster in describing a video sample. Accordingly, our approach intentionally avoids normalization on each descriptor, which is in contrast to [1, 2].

Consider the example video of *board trick* in Fig. 1, which consists of title screens at both ends and actual snowboarding segments in the middle. First, there are a set of segment clusters discovered by clustering segment-level features[2], which include clusters such as caption/titles, moving object on smooth background, and cube-shaped large objects[3]. It is worth noting that our framework is general and can be applied to various audio-visual features developed for multimedia videos. Accordingly, clusters with motion or audio patterns can be discovered as well, depending on the characteristics of the underlying features. For example, it can be seen in Fig. 1 that not only scene-based clusters but

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://–enter the whole DOI string from rightsreview form confirmation.

[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]For these results, (uncolored) HoG3D feature [4] is used.
[3]The clusters are manually named *a posteriori* after clustering.

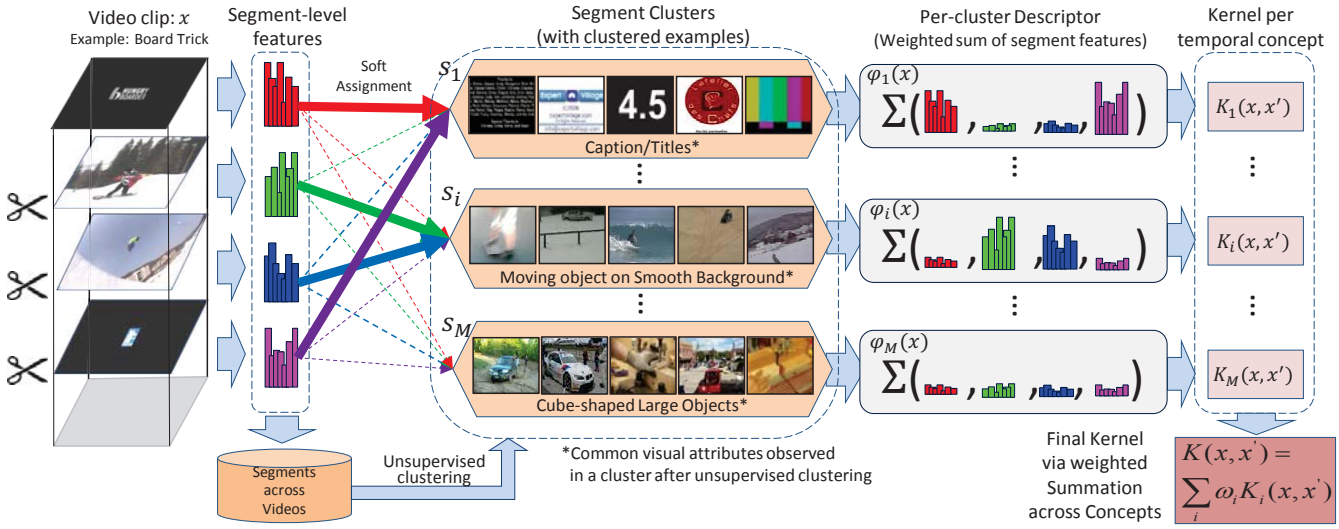


Figure 1: A video clip consists of multiple segments. Each segment-level feature is pooled multi-way into different descriptors based on their similarity and the corresponding segment clusters. Then, kernelization is separately applied per descriptor. The final kernel combines multiple kernels and provides improved discriminant power.

also motion-related clusters are discovered. Then, each per-cluster descriptor is built by pooling features from different parts of the input video using soft-assignment based on similarity between segments and clusters.

For classification, our work adopts the intersection kernel (IK) SVMs to build classifiers. In particular, the kernel between a pair of videos is computed by combining per-cluster kernels computed for every cluster. It is important to note that the per-cluster kernel values on a frequently assigned kernels tend to be high and vice versa, due to the captured frequency information. Accordingly, if certain clusters are poorly represented in exemplar videos, they will contribute towards kernel values in a limited way. In addition, our work explores an alternative strategy to combine kernels using multiple kernel learning (MKL) [11]. In essence, it is plausible that certain clusters are more discriminative even though they are rarely represented in exemplar videos, or vice versa. The use of MKL provides an opportunity to learn discriminative weights for kernel combination.

We have evaluated the proposed approach on the recently introduced challenging TRECVID '11 MED dataset [9]. The experimental results show that the proposed approach shows substantial improvement over the state-of-the-art.

2. RELATED WORK

The idea of multi-way pooling has been developed in [1], but it was only applied to low-level raw visual feature descriptors for image recognition. This work extends it to video recognition at higher-level granularity of segments. Recent work that characterizes videos at segment-level include [7], [10], and [2]. To represent complex activities, [7] identifies distinctive temporal segments (i.e., sub-actions) along with the temporal structure between them. Although the temporal structure allows certain flexibility, [7] is still most suitable to videos with fairly regularized structures (such as the Olympic dataset). In [10], a discriminative recursive hidden segmental Markov model is proposed to cope better with less regularized temporal structure in consumer videos. In the closely related work of [2], features from images in videos are pooled into different scene clusters, guided by the secondary GIST [8] feature. Although using a secondary feature might be necessary to incorporate extremely sparse feature types (e.g., sparse SIFT), it limits this method to be applicable to image-based features only, and exploring a unified feature pooling framework (without a secondary feature) for more general feature types is necessitated. In contrast, our method is simpler and more general because it is more widely applicable to diverse audio-visual features beyond image-based ones, and can utilize audio/temporal clusters. In contrast to [2], our approach also explores MKL variations to combine kernels across different clusters.

3. MULTI-WAY LOCAL POOLING

Our multi-way local pooling (MLP) method constructs multiple descriptors instead of a single descriptor given a feature type for a video clip, and then attempts to improve discriminant power using kernelization techniques. The key idea is to quantify and utilize similarities between two video samples with respect to (w.r.t.) various segment clusters, especially in unconstrained consumer video data, where it is difficult to apply conventional temporal models, e.g., HMMs.

The overall scheme is illustrated in Fig. 1. First, we divide a video clip into video segments and represent each segment with a given feature type. Then, every video segment is soft-assigned to segment clusters. These clusters are pre-constructed by unsupervised clustering from all segment-level feature descriptors in training data, and thus represent broad categories covering entire training corpus, e.g., caption/title, moving object on smooth background, or cube-shaped large object. A large assignment value of a video segment to an existing cluster indicates that they are highly correlated, and vice versa. Soft-assignment is important because it can substantially alleviate the arbitrary space partitioning built by unsupervised clustering of segments. Using this soft-assignment, every segment-level feature descriptor from a video is pooled into multiple different segment clusters with different weights, i.e., multiway-pooling. In other words, if we have M segment clusters, M video-level feature descriptors are constructed in a way that a highly correlated video segment to a corresponding segment cluster contributes more. In a sense, the newly constructed video-level feature descriptors can be considered to be projections

of a video clip toward segment clusters. After multiple descriptors are constructed, kernelization is separately applied to measure similarity between different videos w.r.t. each corresponding segment cluster. Finally, multiple kernels are combined to a final kernel to provide improved discriminant power for video recognition. For brevity, the detailed derivations below are based on BoW features, although it can be generalized to other representations.

In detail, let $x = \{x^i | x^i \in R^D, 1 \leq i \leq n\}$ be a training video sample, where x^i is a D-dimensional BoW representation for the i-th segment, and n is the number of total segments in a video sample x. It is assumed that segment clusters are already available by collecting all of the video segments from the training corpus and clustering them in the D-dimensional feature space by an unsupervised k-means scheme. Then, our approach uses centroids of the clusters as segment clusters that compactly describe the segment types in the video. Let $S = \{s_j | s_j \in R^D, 1 \leq j \leq M\}$ be a set of M segment clusters, which are represented as D-dimensional vectors. Each D-dimensional feature descriptor $\varphi_j(x)$ of a video x w.r.t. the j-th segment cluster is formulated as a weighted-BoW representation computed across the entire video segments $\{x^1, x^2, \cdots, x^n\}$, with corresponding soft-assignment weights as

$$\varphi_j(x) = \frac{1}{n} \sum_{i=1}^{n} \omega_j(S, x^i) \cdot x^i, \tag{1}$$

where n is the number of video segments in a video sample x, and $\omega_j(\cdot)$ is a soft-weight assignment function between a corresponding segment cluster and a video segment. While the choice for the soft-weight assignment is flexible, we have adopted the following variant of the Gaussian function, which has shown superior performance across our experiments with diverse features:

$$\omega_j\left(S, x^i\right) = \exp\left[-\frac{\left\{d\left(s_j, x^i\right)\right\}^2}{\alpha}\right], \tag{2}$$

where α is a positive parameter that controls the sensitivity on the distance $d(\cdot)$ between a centroid and a sample point. For a distance measure, the negative geodesic distance (NGD) which provides an effective distance measure on BoW features [12] is used, defined as the following:

$$d\left(s_j, x^i\right) = -2 \arccos\left(\sum_{k=1}^{D} \sqrt{\frac{s_{j,k}}{|s_j|} \frac{x_k^i}{|x^i|}}\right). \tag{3}$$

For the kernel type, we select IK, which is a popular kernel type, along with the desirable property of not involving normalization during kernel computation. It is noted that the constructed BoW feature $\varphi_j(x)$ for a segment cluster s_j is not L_1-normalized. In this way, the IK $K_j(x, x')$ between two videos x and x' is determined by not only the similarity between the distribution of $\varphi_j(x)$ and $\varphi_j(x')$, but also $\|\varphi_j(x)\|$ and $\|\varphi_j(x')\|$, which reflect their assignment frequency and correlation to a segment cluster s_j. In other words, even if $\varphi_j(x)$ and $\varphi_j(x')$ show similar distributions, $K_j(x, x')$ might be small if one or both samples are loosely correlated to a segment cluster s_j, i.e., $\|\varphi_j(x)\|$ and/or $\|\varphi_j(x')\|$ are small. Then, a final kernel $K(x, x')$ is constructed as a linear combination of the multiple kernels constructed for multiple segment clusters. We applied both equal weights and those learned by MKL, and their results are reported in the following section.

4. EXPERIMENTAL RESULTS

We have applied the proposed framework on TRECVID '11 MED corpus [9], which is a challenging large-scale consumer video dataset. The dataset provides an excellent testbed for a real-world unconstrained video retrieval problem. It consists of 13K training and 32K test samples with 10 annotated test event classes. The number of positive and negative samples is highly imbalanced; e.g., there are only about 150 positive samples for each class in each of the training and test sets. For each event class, we trained SVMs in a one-vs-all manner across our experiments and report average precisions (APs) or mean AP (mAP) across all ten events as metrics. For training protocol, we followed the approach in a recent study [10] for fair comparison, across experiments.

The proposed approach is extensively compared with other strong baseline methods and/or state-of-the-arts for both visual and audio features. For the results in this paper, (visual) HoG3D [4] with 1,000 codewords and (audial) MFCCs with 1,024 codewords are used, respectively. For the length of video segments, we divided a clip uniformly into fixed 2-second-length segments, which we found to work well across feature types. The parameter α in Eq. 2 was set to be $\alpha = 0.3\pi^2$, and was found through cross validation.

Table 1: Retrieval results w.r.t. varying number of segment clusters on HoG3D, in mAP (%).

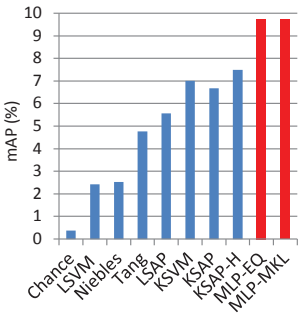
# of SCs	1	20	40	60	80	100
mAP	7.00	9.43	9.75	9.74	9.75	9.72

Our first result analyzes the sensitivity of the proposed framework against the number of segment clusters (SCs). Table 1 summarizes the mAP across all ten classes w.r.t. varying number of segment clusters on HoG3D. It is noted that using only one cluster is equivalent to using a conventional kernelized SVM (KSVM). It can be observed that significant improvement (relatively 39.3%) can be achieved as the number of clusters is increased to 40. Beyond 40, the performance stabilizes, showing the desirable property that the proposed approach is relatively immune to over-fitting even when large number of SCs are used, which can be credited to the proposed distance and soft-weighting schemes. Although the number of optimal segment clusters may differ by features types, event types, and datasets, we find that 30–50 SCs are generally sufficient to acquire the benefits of the proposed framework. For the remainder of the experimental results, 40 clusters have been used.

Our main experimental results on 32K test samples using visual HoG3D features are summarized in Table 2. The performance of Chance (i.e., random) is very low due to the imbalance between positive and negative samples. For MLP approaches, two variations using equal kernel weights (MLP-EQ) and those learned by generalized MKL [11] (MLP-MKL) are reported. The compared approaches used in our experiments include linear SVM (LSVM), KSVM, Niebles [7], Tang [10], and linear/kernelized SAP (LSAP/KSAP) [2]. It is noted that, for direct comparisons, we reproduced the results of Niebles and Tang from [10] by using the same quantized features and training/test protocol, and also re-implemented LSAP/KSAP using the same GIST [8] feature as a secondary image feature and parameters suggested by the authors [2]. In addition, we also evaluated a variant of KSAP (denoted as KSAP-H), in which scene clusters are constructed by using the same pooled feature (not a secondary image feature, suggested by [2]), i.e., HoG3D in this experiment. For KSVM,

Table 2: Comparison in AP(%) among the baseline systems including state-of-the-arts, and the proposed MLP frameworks. For each row, the best result is marked in bold. Overall, both MLP-EQ and MLP-MKL consistently outperformed baselines, showing notable improvement in mAP (illustrated in the right figure).

event ID	Chance	LSVM	Niebles	Tang	LSAP	KSVM	KSAP	KSAP-H	MLP-EQ	MLP-MKL
E006	0.54	1.97	2.25	4.38	3.95	6.08	4.24	4.73	6.34	**6.74**
E007	0.35	1.25	0.76	0.92	2.88	2.87	2.86	2.26	**3.01**	2.98
E008	0.42	6.48	8.30	15.29	17.31	20.75	22.33	22.99	**31.16**	30.87
E009	0.26	2.15	1.95	2.04	4.33	6.25	5.36	**7.61**	7.54	7.50
E010	0.25	0.81	0.74	0.74	1.31	1.43	1.14	1.34	2.11	**2.34**
E011	0.43	1.10	1.48	0.84	1.94	2.29	2.57	2.65	**4.07**	3.86
E012	0.58	5.83	2.65	4.03	7.43	8.44	7.08	8.7	10.63	**11.13**
E013	0.32	2.58	2.05	3.04	9.78	9.44	9.33	10.43	**15.57**	15.25
E014	0.27	1.18	4.39	10.88	5.25	10.00	9.79	11.89	14.81	**14.84**
E015	0.26	0.92	0.61	**5.48**	1.54	2.49	2.02	2.4	2.25	1.82
mAP	0.37	2.43	2.52	4.76	5.57	7.00	6.67	7.50	**9.75**	9.73



KSAP, and KSAP-H, we applied the same IK used in the proposed approach. Across all compared methods, the same HoG3D features have been used.

As shown in Table 2, the proposed approach consistently outperforms the compared systems for most event classes. In particular, the proposed MLP approaches show significant improvement of (relatively) 30% on average beyond KSAP-H, which is found to be the best baseline system. Such improvement was achieved by our soft-assignment scheme based on NGD distance and kernel combination without cluster-wise normalization. We also observed that KSAP-H outperforms KSAP. This implies that the use of a consistent feature in constructing SCs can improve the quality of SCs, when compared to the use of a secondary image feature, especially for densely extracted feature types (we note that our HoG3D feature is densely extracted, while features used in [2] are sparsely extracted). In addition, it can be observed that KSVM outperformed the other latest methods without kernelization (Niebles, Tang, and LSAP), which suggests that the use of kernelization is one of the critical techniques for successful event detection.

Table 3: Results in mAP(%) using MFCCs (MLP with audio features).

	Chance	LSVM	KSVM	MLP-EQ	MLP-MKL
mAP	0.36	1.25	5.98	6.83	**6.87**

In Table 3, we also compared the proposed framework using audio MFCC features against two baselines (LSVM and KSVM). Other baseline systems are not included because the non-image features can not be incorporated or they did not show better performance than KSVM. In this experiment, we excluded about 2% of test videos without audio, which makes the performance by Chance slightly different from Table 2. It is clear that our framework provides advantages in audio as well, showing on average 14.9% improvement (relative) over KSVM. These results show that the proposed framework is fairly general and can yield benefits across different feature modalities.

Among our methods, MLP-EQ and MLP-MKL showed comparable results across all of the event classes, with slight improvement by one or the other, depending on event classes. The surprising effectiveness of MLP-EQ can be attributed to the following reasons: (1) individual kernels constructed for corresponding segment clusters are already weighted by the assignment frequency, which seems to capture most discriminative information; and (2) when underlying features are constructed effectively with little redundancy, equally weighted kernels has been shown to have comparable performance to MKL [5]. These results suggest that, for time-sensitive applications, the use of MLP-EQ alone can be a good approach at the loss of minor accuracy for some classes.

5. CONCLUSION

We presented a novel multi-way feature pooling approach to address the problem of complex event detection on unconstrained videos, especially to capture relevant contents effectively from varying contents embedded within temporal structures. For this purpose, the proposed framework constructs multiple descriptors w.r.t. pre-constructed segment clusters. Our extensive experiments on the challenging TRECVID '11 MED dataset demonstrate the usefulness of the proposed framework, showing promising performance against strong baselines and the state-of-the-art.

6. REFERENCES

[1] Boureau, Y.-L., Roux, N. L., Bach, F., Ponce, J., and LeCun, Y. *Ask the locals: multi-way local pooling for image recognition. In* ICCV *(2011).*

[2] Cao, L., Mu, Y., Natsev, A., Chang, S.-F., Hua, G., and Smith, J. R. *Scene Aligned Pooling for Complex Video Recognition. In* ECCV *(2012).*

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

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

[5] Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. *l_p-norm multiple kernel learning.* JMLR *(March 2011).*

[6] Natarajan, P., Wu, S., Vitaladevuni, S. N. P., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., and Natarajan, P. *Multimodal feature fusion for robust event detection in web videos. In* CVPR *(2012).*

[7] Niebles, J. C., Chen, C.-W., and Fei-Fei, L. *Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In* ECCV *(2010).*

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

[9] Over, P., Awad, G., Fiscus, J., and Antonishek, B. *TRECVID 2011–Goals, Tasks, Data, Evaluation Mechanisms and Metrics.*

[10] Tang, K., Fei-Fei, L., and Koller, D. *Learning Latent Temporal Structure for Complex Event Detection. In* CVPR *(2012).*

[11] Varma, M., and Babu, B. R. *More generality in efficient multiple kernel learning. In* ICML *(2009).*

[12] Zhang, D., Chen, X., and Lee, W. S. *Text Classification with Kernels on the Multinomial Manifold. In* SIGIR *(2005).*