A COMPACT SHOT REPRESENTATION FOR VIDEO SEMANTIC INDEXING

Jinzhuo Wang, Wenmin Wang, Ronggang Wang, Wen Gao*†

School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
†National Engineering Laboratory for Video Technology, Peking University
wangjz@sz.pku.edu.cn, wangwm@ece.pku.edu.cn, rgwang@pkusz.edu.cn, †wgao@pku.edu.cn

ABSTRACT

This paper presents a compact shot representation for video semantic indexing (SIN). The proposed representation consists of visual cues from only two frames, i.e., key frame (KF) and difference frame (DF), which are both constructed with spatial pyramid. The KF describes static information while the generated DF captures non-static information. Each region of DF is derived from the same location in a selected frame, which has the most salient difference compared with the key frame in that region. We introduce a variation of DF to further enhance our model. Experimental results on TRECVID SIN demonstrate that our method obtains better accuracy than the state-of-the-art, while requiring less storage space and consuming time.

Index Terms— Video Semantic Indexing, Compact Shot Representation, Key frame (KF), Difference frame (DF)

1. INTRODUCTION

Video semantic indexing (SIN), aiming at assigning semantic concepts to video shots, is a challenging problem and first presented as an open task in the TRECVID [1]. As an extension from static image classification to video domain, SIN has to cope with a much larger feature space of multi frames which makes the famous semantic gap between low-level features and high-level semantic concepts even more difficult.

Most existing approaches prefer to generate a shot representation which is fed to SVM for classification. One major challenge is to cope with multi-frame visual features since a shot often contains dozens or even hundreds of frames, which has a lot of redundancy [2]. Recently, local features such as HOG/HOF [3], HOG3D [4] and 3D-SIFT [5] based on space-time interest point (STIP) [6] and dense trajectories [7, 8] are used to achieve powerful representations for video sequence, and show good performance for action recognition. But it may fail to be used in general videos such TRECVID SIN dataset which consists of Internet videos from a wide range of semantic concepts such as objects, scene, behavior and event. The reported state-of-the-art approach [9] uses tree-structured Gaussian mixture model (GMM) for visual features densely sampled from the key frame and sparsely sampled from others in a shot, with combination of audio features for SIN exploration. However, their overall model is heavy with a lot of burdens on storage space and consuming time.

This project was supported by Shenzhen Peacock Plan (20130408-183003656).

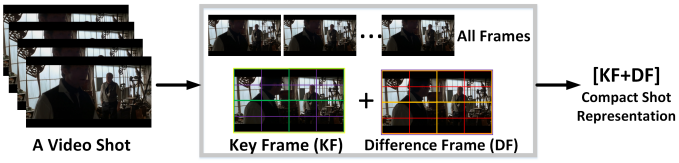

Fig. 1. The generation of our compact shot representation.

In this paper, we propose a compact shot representation leveraging visual features from only two frames. One is the selected key frame to describe static information and the other is a generated one which we call difference frame to capture movement information. We concatenate the visual features of key frame and difference frame to obtain a compact representation for each shot with static and non-static understanding. A variation of difference frame is introduced to further enhance our method. We evaluated our method on TRECVID 2010 SIN benchmark and show promising results.

The remainder of the paper is organized as follows. Section 2 reviews the related work for SIN, and Section 3 proposes our framework with focus on difference frame generation. Experimental results with analysis and comparison are presented in Section 4, and we conclude the paper in Section 5.

2. RELATED WORK

A basic approach for semantic indexing is based on bag-of-feature (BoF) model which classifiers video shot by creating histograms of quantized low-level features. To this end, various attempts have been made from the following two directions, i.e., using only key frame(s) and all the frames in a shot.

For the first direction, [10] employs only the selected key frames (average 7 for one shot) and regards SIN as a multiple instance learning problem, although with the help of advanced locality-constrained linear coding (LLC) strategy, their performance is still below the average. [11] combines the

Fig. 2. Illustrations of generating difference frame. All the frames in a shot are shown at the top. We select D pyramid segments from D frames to form difference frame. Each selected segment has the fewest matches to key frame in the same region.

strength of object detection by region of interest (ROI) detection to explore semantic indexing. [12] leverages shot-based low-level features and early fusion strategy. [13] uses cross-domain fusion. Similarly, [14] provides multi-modal analysis with textual metadata information.

On the other hand, [2] uses visual features and audio features for all the frames in a shot, with GMM and hidden Markov model (HMM) to achieve a uniform representation. [9] proposes a fast solution to the parameter estimation of GMM which uses adaptive maximum a posteriori (MAP), obtaining state-of-the-art performance.

However, both of the two class methods have their own limitations. The first class fails to capture movement information while the second suffers a lot of computation burdens. To overcome these limitations, we attempt to generate a difference frame for each shot to balance the richness and redundancy which can capture the non-static region in different spatial scales. With combination of key frame features, we exploit video semantic indexing with a compact shot representation consisting of visual features from only two frames.

3. APPROACH

We assume that each video is automatically segmented into shots with a selected key frame as provided by TRECVID SIN task. We use spatial pyramid to partition all the frames and generate a difference frame for each shot which is described in Section 3.1. We investigate a variation of difference frame as presented in Section 3.2 and finally we concatenate features of key frame and difference frame to obtain the shot representation for SVM classification in Section 3.3.

3.1. Difference Frame

We show how to generate the difference frame for each shot. First, we describe multi-resolution feature matching between two sets of features, and then employ this technique to select each spatial pyramid segment of difference frame, by searching all the frames in a shot, to find the optimal region with the most salient difference compared with key frame (Fig.2).

3.1.1. Multi-resolution feature matching

The multi-resolution feature matching is first presented in [15] which is able to find an approximate correspondence between feature sets. Let $\mathcal{X}$ and $\mathcal{Y}$ be the two sets of d-dimensional features extracted from two frames in a shot. We place a sequence of increasingly coarser grids over the feature space and take a weighted sum of the number of matches that occur at each level of resolution. At any fixed resolution, two points are thought to match if they fall into the same cell of the grid, and matches found at finer resolutions are weighted more highly than matches found at coarser resolutions. Specifically, a sequence of grids at resolution $0, ..., \mathcal{L}$ is constructed such that the grid at level ℓ has 2^ℓ cells along each dimension, for a total of $D = 2^{d\ell}$ cells. Let $H^\ell_\mathcal{X}$ and $H^\ell_\mathcal{Y}$ denote the histogram of X and Y at this resolution, so that $H^\ell_\mathcal{X}(i)$ and $H^\ell_\mathcal{Y}(i)$ are the numbers of points from $\mathcal{X}$ and $\mathcal{Y}$ that fall into the i-th cell of the grid. Then the number of matches at level ℓ is given by the histogram intersection function [16].

$$\mathcal{I}(H^\ell_\mathcal{X}, H^\ell_\mathcal{Y}) = \sum_{i=1}^{D} \min\left(H^\ell_\mathcal{X}(i) - H^\ell_\mathcal{Y}(i)\right) \qquad (1)$$

Since the number of matches found at level ℓ also includes all the matches found at the finer level $\ell + 1$, the number of new matches at level ℓ is given by $\mathcal{I}^\ell - \mathcal{I}^{\ell+1}$. The weight associated with level ℓ is set to $\frac{1}{2^{\mathcal{L}-\ell}}$, which is inversely proportional to cell width at that level. Intuitively, we want to penalize matches found in larger cells because they involve increasing dissimilar features. Putting all the pieces together, we get the total matches between two feature sets which is

called multi-resolution feature correspondence:

$$\begin{aligned}\mathcal{C}(\mathcal{X},\mathcal{Y}) &= \mathcal{I}^{\mathcal{L}} + \sum_{\ell=0}^{\mathcal{L}-1} \frac{1}{2^{\mathcal{L}-\ell}}(\mathcal{I}^{\ell} - \mathcal{I}^{\ell+1}) \\ &= \frac{1}{2^{\mathcal{L}}}\mathcal{I}^{0} + \sum_{\ell=1}^{\mathcal{L}} \frac{1}{2^{\mathcal{L}-\ell+1}}\mathcal{I}^{\ell}\end{aligned} \tag{2}$$

3.1.2. Selecting pyramid segments for difference frame

Since we have a feature matching measurement between two feature sets, we introduce our application of this technique to generate a difference frame for each shot. Suppose a shot s contains n frames $\mathcal{F} = \{f_1, ..., f_n\}$ and the key frame is labeled as f_k. We use p level spatial pyramid to partition each frame into $2^p \times 2^p$ segments and develop total $\mathcal{N} = \sum_{i=0}^{p} 2^p$ regions for each frame. Each spatial pyramid region in difference frame comes from a selected frame, in which the feature set in that region has the fewest correspondences between that in key frame. Intuitively, the generated difference frame is expected to contain all the salient changes in different spatial scales for a shot. We achieve this by searching the fewest correspondences between feature sets in each region in all frames and key frame. Then the region i in the difference frame is searched by

$$\mathcal{S}_i = \min_{f \in \mathcal{F}} \mathcal{C}(\mathcal{T}_i^f, \mathcal{T}_i^{f_k}) \tag{3}$$

Where $\mathcal{T}_i^f$ denotes the feature set of frame f in region i and the $\mathcal{T}_i^{f_k}$ is the feature set of the same region in key frame. Fig.2 presents an example of generating a difference frame with 2-level pyramid. Note that the segments in p level do not cover the contents of $p-1$ level since we only use the spatial pyramid representation of difference frame.

3.1.3. Variation of difference frame

Each spatial region of difference frame is expected to contain the most salient change in that location in a shot. However, the selection method in Section 3.2.2 uses right the feature set in the region of the selected frame which focus on the static description after a motion happens. We investigate to employ the difference of two feature sets to indicate the non-static change as a variation. For a feature set $\mathcal{X}$ in a SP region of key frame and the selected corresponding feature set $\mathcal{Y}$, we define a set $\mathcal{D}$, calculated as $\mathcal{D}_i = |\mathcal{X}_i - \mathcal{Y}_i|_1$ where i is the i-th descriptors, to represent the difference between two regions. We refer difference frame generated by the select frames in Section 3.1.2 as selected-DF (s-DF), while the one generated using the set $\mathcal{D}$ is denoted as difference-DF (d-DF).

3.2. Shot representation and classification

By now we have obtained difference frame in a universal spatial partition with the key frame. We then concatenate them into a single vector to construct a signature to characterize a shot, which contains both static and non-static information. Local descriptors are coded with LLC strategy [17] to achieve a codebook and we set the size as $4,000$ which has shown empirically to give good results. We use max pooling in a row-wise manner for spatial pooling. For classification, we use simple linear SVM where the penalty parameters are optimized via 5-fold cross validation on the training data using libSVM implementation [18].

4. EXPERIMENTS

4.1. Dataset & Experimental Settings

Our experiments are conducted on TRECVID 2010 SIN dataset which consists of 400h Internet archive videos. The shot boundaries and key frames are automatically detected and provided. Half of the videos are used for training, and the others are used for testing. The task is to detect 30 semantic concepts including objects, events and scenes which are considered meaningful for video exploitation. The labels for training data are created using a collaborative annotation system [19]. The evaluation measures are mean average precision (Mean AP) which is defined as the mean of APs over all 30 target concepts. APs are given as

$$\mathrm{AP} = \frac{1}{R}\sum_{r=1}^{N} \mathrm{Pr}(r) \times \mathrm{Rel}(r) \tag{4}$$

where R is the number of positive samples, N is the number of testing samples, $\mathrm{Pr}(r)$ is the precision at rank of r, and $\mathrm{Rel}(r)$ takes a value of one if the r-th shot is positive; otherwise, it takes zero. The AP is estimated by using a method called inferred average precision (inf AP) as [20].

As for experimental settings, we mainly describe the features in our experiments since the detail implementation of our shot representation is presented aforementioned.

Visual feature. We extract SIFT and hue histogram with dense sampling for all the frames to generate difference frame [21]. This feature combines both gradient and color information, obtaining a 164-dimensional low-level descriptors consisting of 128-dimension SIFT features and 36-dimension hue histograms. PCA is used to reduce the dimensions to 32. SiftGPU implementation is used for SIFT feature extraction [22].

Audio feature. We use mel-frequency cepstral coefficients (MFCC) audio features as complementary cues. The 38-dimension audio features consists of 12-dimension MFCCs, 12-dimension Δ MFCCs, 12-dimension $\Delta\Delta$ MFCCs, 1-dimension Δ log-power, and 1-dimension $\Delta\Delta$ log-power. Here, "Δ" means the derivative of the feature. We implement MFCC extraction wiht a speech recognition toolkit HTK [23].

Method	KF+s-DF	KF+d-DF	+Audio	GMM [9]	Method	KF+s-DF	KF+d-DF	+Audio	GMM
Concept	Inf AP	Inf AP	Inf AP	Inf AP	Concept	Inf AP	Inf AP	Inf AP	Inf AP
Airplane	0.150	0.162	**0.213**	0.117	Animal	0.139	0.152	**0.194**	0.076
Asian-people	0.014	**0.055**	0.052	0.009	Bicycling	0.105	**0.182**	0.121	0.056
Boat-ship	0.051	0.099	**0.164**	0.084	Bus	0.025	0.084	**0.095**	0.016
Car-racing	0.092	0.148	**0.285**	0.043	Cheering	0.092	0.152	**0.247**	0.051
Cityscape	0.132	**0.197**	0.166	0.179	Classroom	0.078	**0.139**	0.126	0.021
Dancing	0.123	0.252	**0.291**	0.067	Dark-skinned-people	0.162	0.192	0.157	**0.203**
Demonstration-or-protest	0.170	**0.192**	0.185	0.132	Doorway	0.082	0.142	**0.179**	0.098
Explosion-fire	0.126	0.230	**0.268**	0.047	Female-face-closeup	0.150	**0.267**	0.184	0.178
Flowers	0.051	**0.138**	0.125	0.044	Ground-vehicles	0.142	0.182	0.199	**0.206**
Hand	0.067	**0.120**	0.106	0.090	Mountain	0.122	0.138	0.148	**0.164**
Nighttime	0.106	**0.182**	0.133	0.132	Old-people	0.051	**0.136**	0.125	0.063
Running	0.143	**0.269**	0.224	0.077	Singing	0.252	0.259	**0.360**	0.188
Sitting-down	0.009	**0.058**	0.043	0.004	Swimming	0.288	**0.354**	0.302	0.276
Telephones	0.124	0.245	**0.285**	0.018	Throwing	0.103	0.152	**0.164**	0.066
Vehicle	0.182	0.214	**0.272**	0.200	Walking	0.160	**0.264**	0.223	0.143
Mean Inf AP	0.150	0.179	**0.188**	0.102					

Table 1. Performance comparison of Inf APs and Mean Inf APs.

4.2. Results & Comparison

Mean Inf APs. Table 1 summarizes obtained Inf AP and Mean Inf AP for two types of DF, fusion with audio features and the state-of-the-art approach [9] which employs tree-structured GMM for the features of all the frames in a shot. As can be seen, our three methods in the left columns perform better than the state-of-the-art method for the majority of semantic concepts. Specifically, for some action-based cases such as *Dancing* and *Throwing*, d-DF yields better accuracy than s-DF. This can be explained that d-DF acts like a motion detector which can capture more movement information. Audio features strength our model in some concepts concerned with sound such as *Car-racing* and *Singing*. Overall, our best performance outperforms the state-of-the-art [9] by 84.3%.

Complexity. We consider complexity on storage space and consuming time. Since our representation consists of visual features from only two frames, certainly we need less space storage as the reported work in the literature using the least frames, to our best knowledge, is [7] which employs average 7 key frames for each shot. As concerned with consuming time, for fair comparison we only discuss the process after feature extraction before SVM training since we use fewer visual features. [9] spent about 3 seconds to make parameter estimation while ours only need less than 1 second to generate difference frame.

4.3. Analysis and discussion

We first look into our model. We investigate to employ different settings to generate our difference frame with main focus on the resolution pyramid level $\mathcal{L}$ to select each segment and the spatial pyramid level p to obtain frame representation. Table 2 lists detailed performance comparison which indicates that $\mathcal{L} = 5$ along with $p = 4$ yields the best accuracy.

Although performing better than the state-of-the-art [9], the overall accuracy is still low compared with that of human annotation. We discuss this issue as follows. First, the dataset consists of low-quality videos upload by normal people and there are some errors with the automatically shot boundary detection and key frame extraction. Second, the SVM training suffers from the extremely unbalanced training data. For instance, for the concept *Bus*, there are only 31 positive shots with $68,980$ negative shots, and for *Car-Racing*, there are 21 positive samples with $101,343$ negative samples. This poses severe challenges to the training algorithm. Third, although with the help of difference frame, our model still lacks the power to detect some actions not contingent to certain scenes, such as *Sitting-down*.

$\mathcal{L}$ \ p	2	3	4	5	6
2	0.138	0.126	0.132	0.140	0.133
3	0.145	0.149	0.151	0.149	0.135
4	0.162	0.170	0.172	0.166	0.143
5	0.155	0.163	**0.188**	0.154	0.152
6	0.137	0.152	0.158	0.149	0.142

Table 2. Mean Inf APs with different parameter settings for d-DF generation with audio features.

5. CONCLUSION

In this paper, we propose a compact shot representation composed of only key frame and difference frame for video semantic indexing. The key frame provides static description while the difference frame captures valuable motion information. Each spatial region of difference frame is derived from a certain selected frame with the most salient difference compared with key frame in the same region. Our method outperforms the state-of-the-art approach by 84.3% on TRECVID SIN 2010 benchmark, while consuming less storage space and execution time.

6. REFERENCES

[1] Paul Over, G Awad, J Fiscus, B Antonishek, AF Smeaton, W Kraaij, and G Quenot, "Trecvid 2010–an overview of the goals, tasks, data," *Evaluation Mechanisms and Metrics*, 2010.

[2] Nakamasa Inoue, Tatsuhiko Saito, Koichi Shinoda, and Sadaoki Furui, "High-level feature extraction using sift gmms and audio models," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3220–3223.

[3] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[4] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[5] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 357–360.

[6] Matteo Bregonzio, Shaogang Gong, and Tao Xiang, "Recognising action as clouds of space-time interest points," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1948–1955.

[7] 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.

[8] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3551–3558.

[9] Nakamasa Inoue and Koichi Shinoda, "A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors," *Multimedia, IEEE Transactions on*, vol. 14, no. 4, pp. 1196–1205, 2012.

[10] Ming Yang, Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Kai Yu, Mert Dikmen, Liangliang Cao, and Thomas S Huang, "Videos semantic indexing using image classification.," in *TRECVID*, 2010.

[11] Xiangyang Xue, Hong Lu, Renzhong Wei, Lei Cen, Yao Lu, and Yingbin Zheng, "Fudan university at trecvid 2010: Semantic indexing.," in *TRECVID*. Citeseer, 2010.

[12] Chao Chen, Qiusha Zhu, Dianting Liu, Tao Meng, Lin Lin, Mei-Ling Shyu, Yimin Yang, Hsin-Yu Ha, Fausto Fleites, and Shu-Ching Chen, "Florida international university and university of miami trecvid 2010-semantic indexing.," in *TRECVID*, 2010.

[13] Yuan Dong, Kun Tao, Hongliang Bai, Xiaofu Chang, Chengyu Dong, Jiqing Liu, Shan Gao, Jiwei Zhang, Tianxiang Zhou, and Guorui Xiao, "The france telecom orange labs (beijing) video semantic indexing systems-trecvid 2010 notebook paper.," in *TRECVID*. Citeseer, 2010.

[14] Miriam Redi, Bernard Mérialdo, Feng Wang, and Yingbo Li, "Eurecom and ecnu at trecvid 2010: The semantic indexing task.," in *TRECVID*, 2010.

[15] Kristen Grauman and Trevor Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. IEEE, 2005, vol. 2, pp. 1458–1465.

[16] Michael J Swain and Dana H Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[17] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[18] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[19] Stéphane Ayache and Georges Quénot, "Video corpus annotation using active learning," in *Advances in Information Retrieval*, pp. 187–198. Springer, 2008.

[20] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam, "A simple and efficient sampling method for estimating ap and ndcg," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 603–610.

[21] Joost Van De Weijer and Cordelia Schmid, "Coloring local feature extraction," in *Computer Vision–ECCV 2006*, pp. 334–348. Springer, 2006.

[22] Krystian Mikolajczyk and Cordelia Schmid, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004.

[23] SJ Young, G Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, D Valtchev, and PC Woodland, "The htk book," 2013.