NHK STRL at TRECVID 2013: Semantic Indexing

Yoshihiko Kawai † Takahiro Mochizuki † Hideki Sumiyoshi † Masanori Sano †

†Science and Technical Research Laboratories, NHK, 1–10–11 Kinuta, Setagaya–ku, Tokyo, Japan

1 Introduction

With the growing popularity of digital video recording devices and the development of high-capacity recording media, it is no longer unusual for individuals to own large amounts of video content. This trend has generated a pressing need for techniques that can efficiently retrieve desired scenes from such large amounts of video data. Broadcast stations as well yearn for a technology that can efficiently search for needed video from the huge amount of broadcast content accumulated over the years so that video assets can be used more effectively.

To achieve effective video retrieval, it is important to analyze semantic content that indicates what the video is showing instead of superficial features. The objective of the semantic indexing task at TRECVID is to detect semantic content such as objects and events, called concepts, appearing in a video [1, 2]. The most successful approaches for analyzing semantic content is the bag-of-visual-words (BoVW) method [3, 4], which is based on the occurrence frequency of local features such as SIFT [5] and SURF [6]. The effectiveness of the BoVW framework has been verified in many prior studies [7], and various researchers continue to make improvements to the elemental techniques of the BoVW framework. Geometric phrase pooling (GPP) [8] is one of the latest techniques for calculating local features. The GPP can calculate highly robust and expressive features by using an integrated analysis of brightness and edge gradients in local regions. A remaining problem with GPP, however, is that it does not reflect visual features in global regions such as the color or texture in an image.

In this paper, we propose a method for detecting semantic concepts by combining local features based on GPP with global features such as color and texture to calculate feature vectors of key frames. A linear support vector machine (SVM) is used for machine learning. In the experiment section, we report the evaluation results for the semantic indexing tasks ("main" and "paired") of TRECVID 2013 [9].

2 Proposed Method

An overview of the proposed method is shown in Fig. 1. First, the input video is divided into shots, and a key frame is extracted from each shot. Next, local feature vectors and global feature vectors are calculated from the key frame and then integrated by using spatial pyramid matching (SPM) [10] to get a feature vector for the entire frame. Local feature vectors are calculated by using GPP, while global feature vectors are calculated by using the color moments (CM), Haar wavelet texture (HWT), and local binary pattern (LBP) [11]. The calculated feature vector is finally classified by using SVM to determine whether a target concept appears in that shot. The SVM method uses training data labeled with concept names to construct the classifier. The following describes each of the processes in detail.

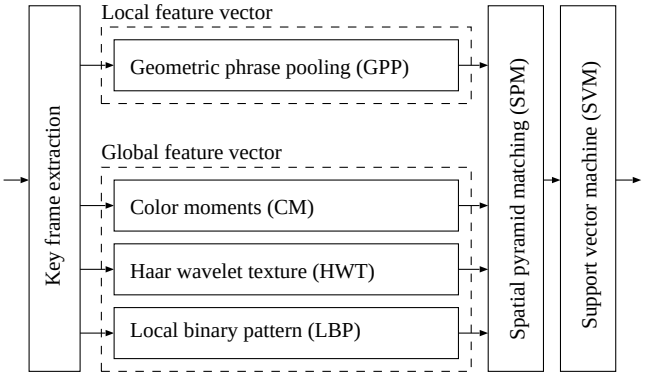

Figure 1: Overview of semantic indexing method

2.1 Local Feature Vector

The procedure for calculating local feature vectors is shown in Fig. 2. SIFT [5] descriptors are first calculated separately for both an input image and an edge-detection image of that input image. The feature points are extracted by dense sampling. The calculated SIFT set $\mathcal{M}$ is expressed as

$$\mathcal{M} = \{(\mathbf{d}_1, \mathbf{I}_1), ..., (\mathbf{d}_M, \mathbf{I}_M)\}. \quad (1)$$

Here, $\mathbf{d}_m$ and $\mathbf{I}_m$ are the descriptor and coordinate, respectively, of the mth feature point, and M is the total number of feature points. Xie *et al.* [8] used the compass operator for edge detection, but the proposed method uses the Sobel operator, taking into account the results of a preliminary experiment.

The SIFT descriptor $\mathbf{d}_m$ is then converted to the B-dimensional coded vector $\mathbf{v}_m$ by using the locality-constrained linear coding (LLC) method [12]. Here, B

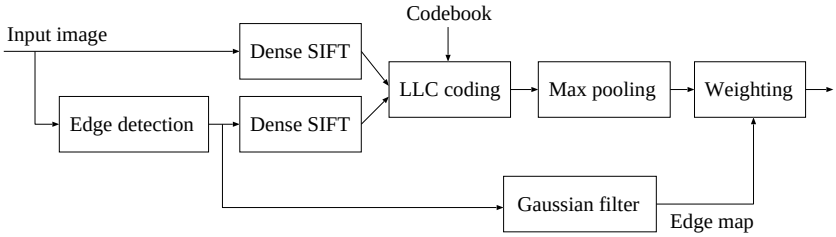

Figure 2: Calculation of GPP local feature vector

denotes the size of the codebook, which is prepared by clustering the SIFT descriptors calculated from training images by using the k-means method and calculating a gravity center for each cluster.

For each feature point, $\mathbf{I}_m$, the proposed method searches for K-nearest neighbor points $\mathbf{I}_{m,k}$ ($k = 1, ..., K$) and integrates a corresponding coded vector, $\mathbf{v}_{m,k}$, by max pooling to create a pooled vector, $\mathbf{w}_m$, taking the peripheral area into account. The formula for calculating $\mathbf{w}_m$ is given as:

$$\mathbf{w}_m = \max_{1 \le k \le K} \{\mathbf{v}_m + s_k \cdot \mathbf{v}_{m,k}\}. \tag{2}$$

The max function is element-wise maximization on K vectors with B dimensions. We set $K = 20$. The parameter s_k denotes a weight based on the distance from feature point $\mathbf{I}_m$ to $\mathbf{I}_{m,k}$. It is defined as

$$s_k = \exp\{-\sigma_w \times ||\mathbf{I}_m - \mathbf{I}_{m,k}||_2\}, \tag{3}$$

where σ_w is a parameter for adjusting the weight. We set $\sigma_w = 0.01$.

Finally, pooled vectors $\mathbf{w}_m$ for each image area are tabulated, and a local feature vector for that area is calculated. The following gives the formula for calculating local feature vector $\mathbf{w}$ when the image area is equal to the entire image.

$$\mathbf{w} = \max_{1 \le m \le M} \{g_m \cdot \mathbf{w}_m\}. \tag{4}$$

Here, g_m denotes a weight for pooled vector $\mathbf{w}_m$. It is defined on the basis of the edge intensity at the feature-point coordinate $\mathbf{I}_m$. The edge intensity is determined by using an edge map created by applying a Gaussian filter to the edge-detection image. The weight of a feature point in an area with a high concentration of edges (with a high possibility of something appearing) is relatively large, while the weight of a feature point in an uneventful area with no edges is relatively small.

2.2 Global Feature Vector

The proposed method uses three types of feature to construct global feature vector.

2.2.1 Color Moments

The proposed method transforms the input image into the HSV and $L^*a^*b^*$ color spaces and then calculates the average pixel value μ_c, the standard deviation σ_c, and the cube root of skewness s_c for each component c ($c \in \{h, s, v, l, a, b\}$). The calculation formulae are as follows:

$$\mu_c = \frac{1}{HW} \sum_x \sum_y f_c(x, y), \tag{5}$$

$$\sigma_c = \left\{ \frac{1}{HW} \sum_x \sum_y \{f_c(x, y) - \mu_c\}^2 \right\}^{1/2}, \tag{6}$$

$$s_c = \left\{ \frac{1}{HW} \sum_x \sum_y \{f_c(x, y) - \mu_c\}^3 \right\}^{1/3}, \tag{7}$$

where $f_c(x, y)$ represents the pixel value of a component c at coordinates (x, y) and H and W are the height and width of the image area.

2.2.2 Haar Wavelet Texture

The proposed method performs Haar wavelet transform of three levels to the image area. We then calculate the variance of the pixel values in each sub-band region, and concatenate them.

2.2.3 Local Binary Pattern

The local binary pattern $L_{P,R}$ from P pixels on a circle of radius R is formulated as

$$L_{P,R}(x, y) = \begin{cases} \sum_{p=0}^{P-1} \delta_{P,R}(x_p, y_p), & if\ U_{P,R}(x, y) \le 2 \\ P + 1, & otherwise \end{cases}. \tag{8}$$

Here, $\delta_{P,R}$ represents the magnitude relationship of intensity values between a particular pixel (x, y) and the surrounding pixels $(x + x_p, y + y_p)$ and is calculated as

$$\delta_{P,R}(x_p, y_p) = \begin{cases} 1, & f(x+x_p, y+y_p) - f(x, y) \ge 0 \\ 0, & otherwise \end{cases}. \tag{9}$$

The values of x_p and y_p are given by

$$\begin{cases} x_p = R \cos \dfrac{2\pi p}{P} \\ y_p = R \sin \dfrac{2\pi p}{P} \end{cases} \quad (0 \le p \le P - 1)\ . \tag{10}$$

Table 1: Evaluation results

Task	Run	System ID	Training type	Local feature	Global feature	Mean infAP
Main	1	NHKSTRL1	A	GPP	CM, HWT, LBP	0.128
	2	NHKSTRL2	A	GPP	—	0.056
Paired	5	NHKSTRL5	A	GPP	CM+HWT+LBP	0.103
	6	NHKSTRL6	A	GPP	—	0.044

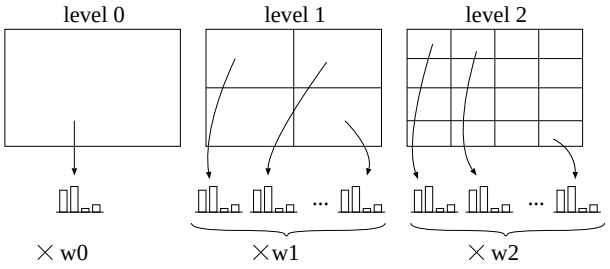

Figure 3: Spatial pyramid matching

The function $U_{P,R}$ in Equation (8) represents the total number of locations where there is a change between 0 and 1 in the sequence $\delta_{P,R}$ for the surrounding pixels, and is calculated by

$$U_{P,R}(x,y) = |\delta_{P,R}(x_{P-1}, y_{P-1}) - \delta_{P,R}(x_0, y_0)| + \sum_{p=1}^{P-1} |\delta_{P,R}(x_p, y_p) - \delta_{P,R}(x_{p-1}, y_{p-1})|. \quad (11)$$

The proposed method calculates $L_{P,R}(0 \leq L_{P,R} \leq P+1)$ for all the pixels in the image area and obtains their frequency histogram. To ensure robustness against changes of resolution, frequency histograms are calculated for each $L_{P,R}$ with $(P,R) = (8,1)$, $(16,2)$, and $(24,3)$ [11].

2.3 Calculation of Feature Vector

The final step is to calculate the feature vector for an entire key frame from the local and global feature vectors. To reflect the spatial, positional information of feature points in feature vectors, the proposed method adopts the spatial pyramid matching (SPM) [10] method. An overview of a feature vector calculated by SPM is shown in Fig. 3. The method uses three-level SPM. The procedure is to tabulate the local and global feature vectors for each area and to then concatenate these to construct a feature vector for the entire key frame.

3 Experiments

For semantic indexing, we submitted two runs for the main task and two runs for the paired task for a total of four runs. Training type A was used in all cases, and the results of past collaborative annotation [13] available from the TRECVID website were used as annotation data for training purposes. Table 1 shows the list of our runs.

Table 2: Evaluation results for each concept (Main)

Concept		infAP	
		Run 1	Run 2
3	Airplane	0.038	0.008
5	Anchorperson	0.548	0.374
6	Animal	0.024	0.008
10	Beach	0.298	0.064
15	Boat_Ship	0.175	0.030
16	Boy	0.006	0.001
17	Bridges	0.019	0.001
19	Bus	0.010	0.001
25	Chair	0.102	0.048
31	Computers	0.061	0.021
38	Dancing	0.127	0.036
49	Explosion_Fire	0.018	0.003
52	Female-Human-Face-Closeup	0.195	0.096
53	Flowers	0.064	0.011
54	Girl	0.016	0.001
56	Government-Leader	0.220	0.093
59	Hand	0.114	0.051
71	Instrumental_Musician	0.328	0.198
72	Kitchen	0.012	0.001
80	Motorcycle	0.145	0.006
83	News_Studio	0.621	0.408
86	Old_People	0.215	0.119
89	People_Marching	0.024	0.004
100	Running	0.053	0.004
105	Singing	0.069	0.042
107	Sitting_Down	0.000	0.000
117	Telephones	0.005	0.000
120	Throwing	0.062	0.001
163	Baby	0.071	0.005
227	Door_Opening	0.012	0.001
254	Fields	0.050	0.006
261	Flags	0.092	0.041
267	Forest	0.040	0.002
274	George_Bush	0.299	0.050
342	Military_Airplane	0.021	0.004
392	Quadruped	0.014	0.001
431	Skating	0.198	0.193
454	Studio_With_Anchorperson	0.496	0.208
	Mean infAP	0.128	0.056

3.1 Main task

The main task is to detect 60 types of single concepts from about 200 hours of test video. We performed two runs of Run 1 and Run 2 for the main task, as shown in Table 1. These two runs differed in the method of calculating features. In Run 1, we calculated feature vectors by using a local feature vector based on GPP and global feature vector based on three types of features, while in Run 2, we used only the local feature vector.

In the evaluation experiment, mean infAP values of

0.128 and 0.056 were obtained for Run 1 and Run 2, respectively. These results demonstrate that the detection accuracy of semantic concepts increases significantly by using the global feature vector in addition to the local feature vector.

Evaluation results for specific concepts are listed in Table 2. As shown, accuracy was higher in Run 1, which used both local and global feature vectors for almost all concepts. An exceptionally high accuracy of approximately 0.5 was achieved in Run 1 for the concepts of "5 Anchorperson" and "454 Studio_With_Anchorperson." We consider that a classifier constructed from training data should also function well with test data because the pattern formed by an anchorperson in news video is very nearly fixed. A high accuracy of approximately 0.3 was also obtained in Run 1 for the concepts of "10 Beach," "71 Instrumental_Musician," and "274 George_Bush." In these cases as well, we consider that video patterns that are essentially fixed made it easy to capture video characteristics.

In comparison, the evaluation revealed low accuracy for the concepts of "16 Boy" and "54 Girl." It can be said that judging a person's age solely on the basis of the image features used in this study was difficult. Additionally, as the proposed method does not take movement information into account, classifying an action as that in "107 Sitting_Down" was difficult. To improve accuracy, we will need to study the combined use of various types of features, including audio features and temporal features.

3.2 Paired task

The paired task is to detect 10 types of concept pairs from about 200 hours of test video. The proposed method trained a classifier for each of the two concepts, making up a concept pair, and then, for each shot, applied these two classifiers and computed the mean of their resulting scores. We submitted Run 5 and Run 6 for the paired task, as shown in Table 1. As in the main task, these two runs differed in the method of calculating features.

Table 1 showed that Run 5, using both local and global feature vectors, achieved higher accuracy than did Run 6, which used only the local feature vector. We consider that taking global features such as color and texture into account improved the accuracy of classifying each concept, thereby improving the accuracy of classifying a concept pair.

The evaluation results for each concept are listed in Table 3. The overall level of accuracy was much lower than that of the results for the main task. Except for a few concept pairs like "917 Chair + George_Bush," accuracy was under 1%, which indicates the need for enhancements to the proposed method. For example, instead of computing the mean of individual classifier outputs in subsequent processing, a better approach may be to create training data, especially for concept pairs and train classifiers, accordingly. We aim to study such alternative methods in the future.

Table 3: Evaluation results for each concept (Paired)

	Concept	infAP	
		Run 5	Run 6
911	Telephones + Girl	0.000	0.001
912	Kitchen + Boy	0.007	0.001
913	Flags + Boat_Ship	0.116	0.042
914	Boat_Ship + Bridges	0.002	0.032
915	Quadruped + Hand	0.000	0.005
916	Motorcycle + Bus	0.000	0.001
917	Chair + George_Bush	0.505	0.097
918	Flowers + Animal	0.000	0.000
919	Explosion_Fire + Dancing	0.000	0.000
920	Government-Leader + Flags	0.395	0.259
	Mean infAP	0.103	0.044

4 Conclusion

We proposed a method for detecting semantic concepts appearing in video streams by combining local features based on geometric phrase pooling and global features such as color and texture that consider larger areas. The evaluation result for semantic indexing showed that the proposed method obtained mean infAP values of 0.128 and 0.103 in the main task and paired task, respectively. In future research, we plan to study the use of temporal information and audio features to further improve the detection accuracy.

References

[1] A.F. Smeaton, P. Over and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. ACM MIR'06*, pp.321–330, 2006.

[2] A.F. Smeaton, P. Over and W. Kraaij, "High-level feature detection from video in TRECVid: a 5-year retrospective of achievements," *Multimedia Content Analysis, Theory and Applications*, pp.151–174, 2009.

[3] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proc. ICCV'03*, 2003.

[4] G. Csurka, C.R. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, pp.59–74, 2004.

[5] D.G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, vol.2. pp.1150–1157, 1999.

[6] H. Bay, T. Tuytelaars and L.V. Gool, "SURF: speeded up robust features," *Computer Vision and Image Understanding*, vol.110, no.3, pp.346–359, 2008.

[7] "TRECVID notebook papers and slides," http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

[8] L. Xie, Q. Tian and B. Zhang, "Spatial pooling of heterogeneous features for image applications," in *Proc. ACM Multimedia*, pp.539–548, 2012.

[9] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A.F. Smeaton and G. Quenot, "TRECVID 2013 – An overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TRECVID 2013*, 2013.

[10] S. Lazebnik, C. Schmid and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE CVPR'06*, pp.2169–2178, 2006.

[11] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. PAMI*, vol.24, no.7, pp.971–987, 2002.

[12] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE CVPR'10*, pp.3360–3367, 2010.

[13] S. Ayache and G. Quenot, "Video corpus annotation using active learning," in *Proc. ECIR'08*, pp.187–198, 2008.