Ritsu_CBVR at TRECVID-2010

Ai Danni, Okamoto Atsusi, Yae Kikutani, Tanaka Yoshiyuki, Xianhua Han, Yen-wei Chen
Graduate School of Science and Enginnering, Ritsumeikan University, Japan

Abstraction

In this paper, we describe our first participation for the semantic indexing task at TRECVID 2010 [1]. We focus on extraction multiple low-level feature sets and a fusion method. In our system, six features are extracted for all the predefined concepts from the keyframes, including global features (RGB color histogram, HSV color histogram, edge histogram, Grey Level Co-occurrence Matrix, GIST) and a local feature (gray-scale SIFT). SVM-based classifiers are trained by utilizing these features and multiple feature weighted fusion of the classification results are used as a baseline.

In this year, only one run was submitted to "full" submission:

F_A_IIPLA_Ritsu_CBVR_1: Multiple feature weighted fusion of classification results based on global features and local features are utilized. SVM classifiers are trained on the images provided by the collaborative annotation in TRECVID 2010.

1. Overview

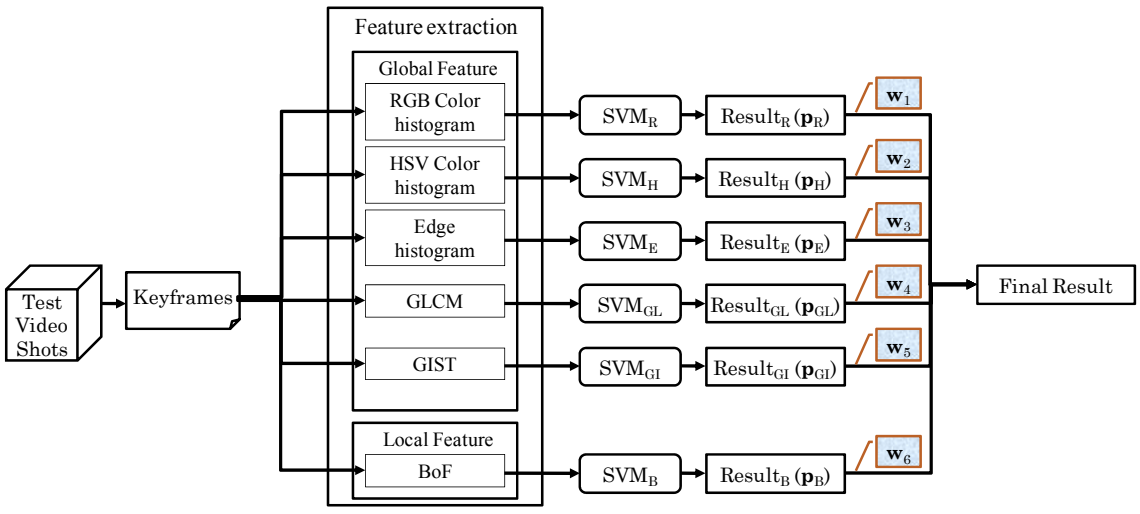

Fig. 1 An overview of Ritsu_CBVR system

An overview of our system is shown in Fig.1. Local features and global features are extracted from each keyframe, which is considered as representative images of shots. For our first participation to the TRECVID semantic indexing task, a very limited set of visual descriptors were available: five global features and one local feature are extracted from each keyframe per shot. SVMs are utilized as classify. Subsequently, multiply feature weighted fusion method is utilized for the semantic indexing. Here, all the positive annotation keyframes provided by the collaborative annotation effort [6] are directly used as training data, which are utilized to train the SVMs classifies and construct the weighting vector $\mathbf{w}_i$ (i=1,2,…,6).

2. Visual Representation

Basically, the image features proposed by researchers fall into two categories: global features calculated over the entire image and local features computed over a small group of pixels [16]. In our current system, six features are extracted and will be introduced in detail

2.1 Global Features

(1) RGB Color Histogram

RGB color histogram [2] is a combination of three histograms based on the R, G and B channels in the RGB color space. We extracted 512-dimensional RGB color histogram feature from each image.

(2) HSV Color Histogram

HSV color histogram [3] is given to calculate the histogram over three channels of HSV color space. Since the three channels of HSV color space are not correlated and HSV color space is similar to the human cognitive system, HSV color space could give more information than RGB color space. The transformation of HSV is shown in the following equations.

$$
\begin{aligned}
H &= \begin{cases} 0, & \text{if} \quad \max = \min \\ \left(60^\circ \times \dfrac{G-B}{\max - \min} + 360^\circ\right) \bmod 360^\circ, & \text{if} \quad \max = R \\ 60^\circ \times \dfrac{B-R}{\max - \min} + 120^\circ, & \text{if} \quad \max = G \\ 60^\circ \times \dfrac{R-G}{\max - \min} + 240^\circ, & \text{if} \quad \max = B \end{cases} \\
S &= \begin{cases} 0, & \text{if} \quad \max = 0 \\ \dfrac{\max - \min}{\max} = 1 - \dfrac{\min}{\max}, & \text{otherwise} \end{cases} \\
V &= \max
\end{aligned} \tag{1.1}
$$

We extracted 500-dimensional HSV color histogram feature from each image.

(3) Edge Histogram

Edge histogram [9] captures the spatial distribution of edges. A given image is first divided into 4×4 non-overlapping blocks. For each block, edges are broadly grouped into seven categories: vertical, horizontal, 45ºdiagonal, 135ºdiagonal and isotropic (nonorientation specific). Therefore, the histogram of each block represents the relative frequency of occurrence of the 5 types of edges in the corresponding block. As a result, each local histogram contains 5 bins. 80 histogram bins are obtained because of 16 blocks in a image.

(4) Gray Level Co-occurrence Matrix (GLCM)

GLCM, proposed in [7], captures the spatial relation at several scales and orientations. It is one of the most known texture analysis methods, and estimates image properties related to second-order statistics. Each entry x(i,j) in GLCM corresponds to the number of occurrences of the pair of gray levels i and j which are a distance d apart in original image [8]. A set of gray-scale spatial dependence probability distribution matrices for a given image block is computed, and 14 textural features which can be extracted from each of these matrices are suggested. To reduce the computational complexity, in particular, we use the computed variance at each scale and orientation in order to index the texture information. Our experiment employs 16 bins for a gray-level quantization.

(5) GIST

The GIST feature can encode edges and textures information in the original image coarsely [15]. The image is first divided into 4×4 blocks and convolved with Gabor filters at 4 scales and 8

orientations. Therefore, the dimension of GIST feature is 512-D.

2.2 A local feature

(1) Bag-of-Words Feature

In computer vision, local descriptors (i.e. features computed over limited spatial support) have proved well adapted to matching and recognition tasks, as they are robust to partial visibility and clutter. In this paper, we use grid-sampling patches, and then compute appearance-based descriptors on the patches. In contrast to the interest points from the detector, these points can also fall onto very homogeneous areas of the image. After the patches are extracted, the SIFT [4] descriptor is applied to represent the local features. The SIFT descriptor computes a gradient orientation histogram within the support region. For each of 8 orientation planes, the gradient image is sampled over a 4 by 4 grid of locations, thus resulting in a 128-dimensional feature vector for each region. A Gaussian window function is used to assign a weight to the magnitude of each sample point. This makes the descriptor less sensitive to small changes in the position of the support region and puts more emphasis on the gradients that are near the center of the region [4]. These SIFT features are then clustered with a k-means algorithm using the Euclidean distance. Then we discard all information for each patch except its corresponding closest cluster center identifier. For the test data, this identifier is determined by evaluating the Euclidean distance to all cluster centers for each patch. Thus, the clustering assigns a cluster c(x) to 1, ...C to each image patch x and allows us to create histograms of cluster frequencies by counting how many of the extracted patches belong to each of the clusters. The histogram representation h(X) with C bins is then determined by counting and normalization such that:

$$h_c(X) = \frac{1}{L_X} \sum_{l=1}^{L_X} \delta\left(c, c\left(x_l\right)\right) \tag{1.2}$$

where δ denotes the Kronecker delta function. Figure 1 shows the procedure bag-of-words(BoW) feature extraction and the extracted histogram feature of example images [5].

3. Classification

2.1 Support Vector Machines

SVMs, as classifiers, has been widely used and shown to be efficient [10]. In our experiments, the

nonlinear SVM classifiers with χ^2 RGF kernel are employed, which are based on 512-D RGB color histogram, 500-D HSV color histogram, 80-D edge histogram, 256-D GLCM descriptors, 512-D GIST descriptors and 128-D BoF descriptors, separately. For multiclass problem, the one-to-all strategy is applied for resolving binary classification in SVM. Here, LibSVM package has been employed [11].

2.2 Multiple Feature Weighted Fusion Method

How to fusion different types of feature would have great affect for classification(i.e. to decide whether a shot contains a concept or not). Most of state of the art algorithms just simply concatenated different feature together. In this paper, we build a SVM classifier for each feature type, then, we have six classification results from RGB color histogram, HSV color histogram, edge histogram, GLCM descriptors, GIST descriptors and BoF descriptors, respectively. We fusion the six results by weighted combination according to their discriminate properties proposed by Xiaojun Qi et al.[12], which is called multiple feature weighted fusion method. Since different features have different discriminate properties, it is important to use adaptive weights for feature fusion. Our system explored to use the features weighted fusion method for combining six features introduced above. The flowchart of multiple feature weighted method is shown in Fig. 1 [13].

In Fig.1, adaptive weights **w**s are pre-computed according to different features' discriminate properties on validation datasets. Therefore, the final SVM probability is denoted as follows:

$$\mathbf{p} = \mathbf{w}_1 \cdot \mathbf{p}_R + \mathbf{w}_2 \cdot \mathbf{p}_H + \mathbf{w}_3 \cdot \mathbf{p}_E + \mathbf{w}_4 \cdot \mathbf{p}_{GL} + \mathbf{w}_5 \cdot \mathbf{p}_{GI} + \mathbf{w}_6 \cdot \mathbf{p}_B \qquad (1.3)$$

where $\mathbf{p}_R$, $\mathbf{p}_H$, $\mathbf{p}_E$, $\mathbf{p}_R$, $\mathbf{p}_{GL}$, $\mathbf{p}_R$, are denoted as the probability vector obtained from RGB-based SVMs, HSV-based SVMs, edge-based SVMs, GLCM-based SVMs, GIST-based SVMs and BoF-based SVMs, respectively; "·" denotes the inner product operation; $\mathbf{w}_i$ (i=1,2,…6) determine the contribution from the above SVMs separately and are automatically estimated by applying the likelihood normalization method [14]. The number of components in each above vector equals that of predefined concept (130). Each component in $\mathbf{p}_X$ (X denoted as R, H, E, GL, GI and B) indicated the probability of a image to be classified as each corresponding concept. As a result, each component in **p** indicates the final probability of an image to be classified as each corresponding concept. To obtain the weight vectors $\mathbf{w}_i$ (i=1,2,…6), we first define the weight vector $\mathbf{w}_X$ (X denoted as R, H, E, GL, GI and B), which are based on above features:

$$\mathbf{w}_X = [w_{X,1}, w_{X,2}, \cdots, w_{X,K}]$$
$$w_{X,k} = \frac{(1/NK)\sum_{n=1}^{N}\sum_{c=1}^{K} L_X(n,c)}{(1/N)\sum_{n=1}^{N} L_X(n,k)} \quad (1.4)$$

where N is the number of testing images and K is the number of predefined categories. Each value $L_X(n,c)$ indicates the probability of image n to be classified as category c by using each feature-based SVMs.

Weight vectors $\mathbf{w}_i$ (i=1,2,…6)can be computed as the following:

$$\mathbf{w}_i = \mathbf{w}_X / (\mathbf{w}_R + \mathbf{w}_H + \mathbf{w}_E + \mathbf{w}_{GL} + \mathbf{w}_{GI} + \mathbf{w}_B) \quad (1.5)$$

where " / " denotes the element-wise division operation.

4. Discussion

Since this is the first time for our team to take part in TRECVID and its semantic indexing task, the performance has been subpar and is substantially below expectations. However, we had invaluable experiences, observations and team's cooperation after this competition. For the semantic indexing task, we submitted only one run because of the limited time. The possible reasons for the failed experiment are analyzed as follows:

(1) The keyframes provided by the collaborative annotation effort are rather noisy (keyframes are wrongly annotated or skipped when they are unambiguously defined). Moreover, only positive keyframes, whose numbers are rather unbalance for each annotation, are used for training. These may result in fallacious training results, which make the final results worse. The distributions of positive and negative keyframes in the training sets pose an important problem. To address this issue, preprocess should be carried out for cleaning the annotations and refining more reasonable training images.

(2) The available low-level features may not quite fit to TRECVID database. More efficient features should be combined together for this task.

(3) The original fusion algorithm is not quite efficient for our feature fusion. For the numerous image sets and kinds of features, an improved fusion method should be proposed.

(4) Our current system is very slow and computational complexity, which really jeopardizes its scalability. It can be felt strongly when processing all the video in excess of ten thousands.

Our future work includes use of effective features and expansion of the fusion algorithm to reduce

the computation cost and raise of classification results.

Reference

[1]Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722

[2] M.Swain and D.Ballard, "Color Indexing", Int. Journal of Computer Vision, vol.7, no. 1, pp. 11-32, 1991.

[3] Shamik Sural , Gang Qian , Sakti Pramanik. SEGMENTATION AND HISTOGRAM GENERATION USING THE HSV COLOR SPACE FOR IMAGE RETRIEVAL. International Conference on Image Processing (ICIP). 2002: p. 589-592

[4] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. The International Journal of Computer Vision, 2004.

[5] Xian-Hua Han, Yen-Wei Chen. Image Categorization by Learned PCA Subspace of Combined Visual-words and Low-level Features. 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing.

[6] St´ephane Ayache and Georges Qu´enot. Video corpus annotation using active learning. Advances in Information Retrieval, pages 187–198, 2008.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 3, no. 6, pp. 610–621, 1973.

[8] Mari Partio , Bogdan Cramariuc , Moncef Gabbouj , Ari Visa. Rock Texture Retrieval using Gray Level Co-occurrence Matrix

[9]Park, D. K., Jeon, Y. S., and Won, C. S. 2000. Efficient use of local edge histogram descriptor. In *Proceedings of the 2000 ACM Workshops on Multimedia*, New York, NY, 51-54.

[10]B. Scholkopf and A. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge, MA, 2002.

[11] C. C. Chang and C. J. Lin. LIBSVM: A Library for Support Vector Machines. Software available at: http://wwwcsie.ntu.edu.tw, 2001.

[12] X.Qi, Y.Han, "Incorporating multiple SVMs for automatic image annotation", Pattern Recognition Vol.40, pp.728-741, 2007.

[13] Yae Kikutani, Atsushi Okamoto, Xian-Hua Han, Xiang Ruan and Yen-Wei Chen. Hierarchical Classifier with Multiple Feature Weighted Fusion for Scene Recognition. PRMU.pp.175-179 , 2010

[14] S. Tamura, K. Iwano, S. Furui, A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, pp. 469–472.

[15] A. Oliva and A. Torralba, "Modeling the shape of a scene: a holistic representation of the spatial envelope," *Int'l J. of Comp. Vision*, vol. 42, pp. 145–75, 2001.

[16] R. Datta, D. Joshi, J. Li and J. Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Survers, vol.40, no.2, pp.1-60, 2008.