ISM TRECVID2009
High-level Feature Extraction

Tomoko Matsui[1], Shin'ichi Satoh[2], Yuji Uchiyama[3]

[1]Institute of Statistical Mathematics, Tokyo, Japan
[2]National Institute of Informatics, Tokyo, Japan
[3]Picolab Co., Ltd, Tokyo, Japan

ABSTRACT

We studied two methods for the high-level feature extraction (HLF) task: (1) a method based on support vector machines (SVMs) with walk-based graph kernels [1], and (2) a method based on the prefixspan boosting (pboost) algorithm [2]. In the former method, each image is first segmented into a finite set of homogeneous segments and then represented as a segmentation graph where each vertex is a segment and edges connect adjacent segments. Given a set of features associated with each segment, we then obtain a positive definite kernel between images by comparing walks in the respective segmentation graphs, and image classification is carried out with an SVM based on this kernel. In the latter method, discriminative image subsequence patterns are mined using pboost, and used for image sequence classification. We submitted four runs using the former method with several combinations of the SVM scores with different kernel and SVM parameters, and two runs using the latter method with different amounts of training data.

1. INTRODUCTION

The HLF task can be regarded as a set of supervised binary classification tasks, where each shot must be assigned a set of binary labels to indicate whether or not it belongs to each concept class. Unlike more specific tasks such as face or character recognition, the emphasis in HLF is on obtaining generic and versatile automatic tools that can learn any concept from a set of examples belonging to the concept class.

For the HLF task, we investigate two strategies: (1) a conventional strategy using a keyframe image extracted for each shot and applying a method based on SVMs with walk-based graph kernels, and (2) a strategy using a shot image sequence and applying a method based on pboost.

2. SVM-BASED METHOD WITH WALK-BASED GRAPH KERNELS

In this method, each image is first automatically segmented into a finite set of "homogeneous" segments and then represented as a segmentation graph, where each vertex is a segment and edges connect adjacent segments. A set of features such as size, color, and texture are associated with each segment. Using this graph-based representation, we apply a graph classification method to classify the images. More precisely, we investigate the use of graph kernels in combination with support vector machine (SVM) classification.

Figure 1 shows three steps of our method for HLF: (i) image segmentation, (ii) kernel calculation, and (iii) SVM classification. In (i), each input image is automatically segmented and represented as a segmentation graph, as explained in Section 2.1. In (ii), a walk-based positive definite kernel between

segmentation graphs is computed, as explained in Section 2.2. Finally, HLF treated as a set of binary classification problems is performed with an SVM using the walk-based kernel between segmentation graphs to classify images.

2.1. Graph-based representation of images

The first step of our approach is to automatically split each image into a variable number of homogeneous regions, using an unsupervised segmentation method [3], as in Figure 2. The image is then represented as a *segmentation graph*, i.e., a simple graph $G = (V, E)$, whose vertices V are the segments obtained by automatic segmentation and whose edges E connect vertices corresponding to adjacent segments of the image. The number of vertices (i.e., of segments) depends on the image. Furthermore, each segment is characterized by a set I of 23 features presented in Table 1. The 12 texture features (nos. 12–23) are the responses to a small filter bank of orientation and spatial-frequency selective linear filters [4]. For each segment $v \in V$ of a segmentation graph, we denote by $F(v) = (f_i(v))_{i \in I} \in \mathbf{R}^I$ the vector of the features (23-dimensional in our case).

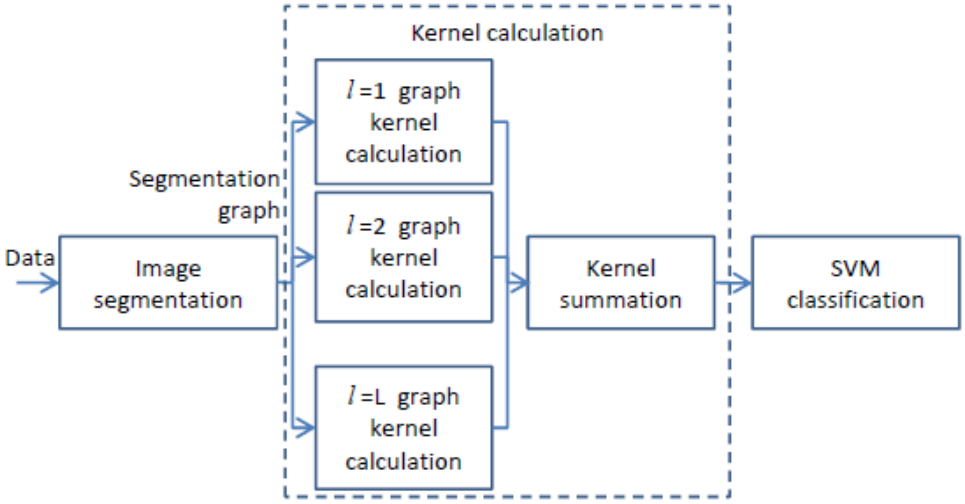

Figure 1. Overall procedure of our method.

Figure 2. Example of segmented image from data set of TRECVID2005.

Table 1. Features characterizing each image segment.

Feature no.	Description
1	Average *x*
2	Average *y*
3	Area in pixels
4	Boundary length divided by area
5	Second moment of area
6–8	Average red, green, blue (RGB) intensities
9–11	Standard deviations of RGB intensities
12–23	Texture features

2.2. Walk-based graph kernel

The walk-based graph kernel is defined hierarchically [5-11]. First, a walk w in a graph $G = (V, E)$ is defined as a finite sequence of connected vertices, i.e., $w = (v_1, ..., v_l)$ with $v_i \in V$ for $i = 1, \ldots, l$ and $(v_i, v_{i+1}) \in E$ for $i = 1, \ldots, l$. Here, l is called the length of walk w. We denote by $W_l(G)$ the set of walks of length l in G.

Second, positive definite kernels between vertices are defined. For any vertices in two graphs $v_1 \in V(G_1)$ and $v_2 \in V(G_2)$, we define a kernel between v_1 and v_2 as a kernel between their respective features, e.g., a Gaussian kernel:

$$K_V(v_1, v_2) = \exp\left(-\gamma \left\| f_I(v_1) - f_I(v_2) \right\|^2\right) = \exp\left(-\gamma \sum_{i \in I} (f_i(v_1) - f_i(v_2))^2\right). \quad (1)$$

Given two walks of length l in two graphs $w = (v_1, ..., v_l) \in W_l(G)$ and $w' = (v'_1, ..., v'_l) \in W_l(G')$, a walk kernel between w and w' is defined as the function:

$$K_W(w, w') = \prod_{i=1}^{l} K_V(v_i, v'_i). \quad (2)$$

Then, the walk-based graph kernel of depth l between two graphs G and G' is defined as

$$K_l(G, G') = \sum_{w \in W_l(G)} \sum_{w' \in W_l(G')} K_W(w, w'). \quad (3)$$

It should be noted that if $l = 1$, no adjacency information is taken into account in the kernel. An image is then considered to be a "bag-of-segments", and the kernel between two images is simply the sum of the vertex kernels between all possible pairs of segments. When $l > 1$, the adjacency information is taken into account.

Finally, we define the walk-based kernel as the sum for multiple depths $l = 1, \ldots, L$ between two graphs G and G' as

$$K_{L-SUM}(G, G') = \sum_{l=1}^{L} K_l(G, G'). \quad (4)$$

We implemented the walk-based graph kernel using a recursive process, as explained in [7]. Since this kernel is positive definite, we can perform image classification with an SVM using the kernel on the segmentation graph representation of the images.

3. PBOOST BASED METHOD

Nozowin et al. [2] proposed a method "pboost" for action classification in videos, using discriminative subsequence mining, which uses the prefixspan algorithm by Pei et al. [12] to find optimal discriminative

subsequence patterns, in combination with linear programming boosting classifier. The idea of boosting classifiers is to combine multiple weak classifiers into a powerful composite classifier. In pboost, the presence of a single discriminative image subsequence pattern in a shot is checked by weak hypotheses, which have the form $h(\boldsymbol{x}; \boldsymbol{s}, \omega)$, where $\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{i = 1, ..., n}$, $\boldsymbol{x}_i$ is a shot image sequence, $\boldsymbol{s} \in S$ is a image subsequence, S is a set of all subsequences of $\{\boldsymbol{x}_i\}_{i = 1, ..., n}$, and $\omega \in \Omega$, $\Omega = \{-1, 1\}$ is an extra variable that allows to decide for either class decision. The classification function has the form:

$$f(x) = \sum_{(s,\omega)\in S\times\Omega} \alpha_{s,\omega} h(x; s, \omega) \tag{5}$$

where $\alpha_{s,\omega}$ is a weight for image subsequence $\boldsymbol{s}$ and parameter ω such that $\sum_{(s,\omega)\in S\times\Omega} \alpha_{s,\omega} = 1$ and $\alpha_{s,\omega} \geq 0$, which indicates the discriminative importance of a image subsequence pattern.

The primal form of the training problem is:

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\xi,\rho} -\rho + D\sum_{i=1}^{n} \xi_i \\ &\text{s.t.:} \sum_{(s,w)\in S\times\Omega} y_i \alpha_{s,\omega} h(x_i; s, \omega) + \xi_i \geq \rho, \\ &\quad \sum_{(s,w)\in S\times\Omega} \alpha_{s,\omega} = 1, \boldsymbol{\alpha} \geq 0, \boldsymbol{\xi} \geq 0 \end{aligned} \tag{6}$$

where $\boldsymbol{y}_i \in \{1,-1\}$ is a class label, ρ is the soft margin separating negative from positive samples, $D = 1/\nu l$ and $\nu \in (0, 1)$ is a hyper-parameter controlling the cost of misclassification. We expanded pboost to deal with unbalanced amount of training samples by using different cost weights for positive and negative samples.

For pboost, a sequence type is constrained to be an ordered sequence of sets of discrete numbers. Here we first clustered all the segments which are denoted as 23-dimensional vectors (as explained in Section 2.1), and then represented each image using a set of cluster IDs which correspond to all segments in the image.

4. DESCRIPTION OF OUR SUBMITTED RUNS

We submitted four runs using SVM based with walk-based graph kernels and two runs using pboost method. For SVM based with walk-based graph kernels, we used 36,262 keyframes in the TRECVID 2008 development data set for training and put the annotations by ourselves. We conducted 3-fold cross validation experiments to find relevant combinations of the values of γ = 8, 16 in eq. (1) and the penalty parameter of the error term C = 0.001, 0.1, 1, 10, 100 in SVM. We set I = [1, 23] and L = 5 for the walk-based graph kernel. The four runs are:

[A_ISM1_1] use of the SVM scores for the best combination of values of γ and C for each feature,

[A_ISM2_2] use of the SVM scores averaged for the top two combinations of values of γ and C for each feature,

[A_ISM3_3] use of the SVM scores averaged for the top three combinations of values of γ and C for each feature, and

[A_ISM4_4] use of the SVM scores for common values of γ=16 and C=1 for all features.

The third run performed the best and the inferred AP was 0.031 on average (figure 3).

For pboost method, we utilized the pboost sequence boosting toolbox[13]. Due to the computation cost and the submission deadline, we selected 5,000 or 10,000 training shots for each feature so as to have almost the same ratio of the numbers of the positive and negative shots for each feature. We also constrained the maximum length of image subsequences to three. The inferred APs with 5,000 training shots are shown in figure 4 and the average AP was 0.017.

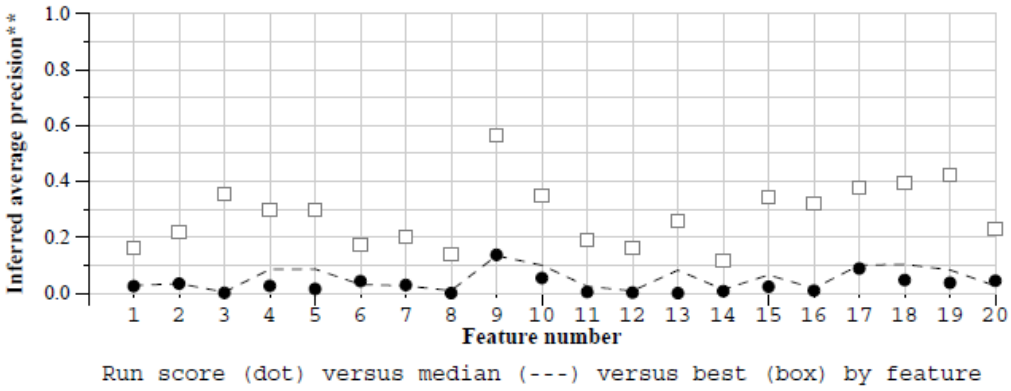

Figure 3. Inferred APs of SVM based method with walk-based graph kernel using the average score for the top two combinations of γ and C values through cross-validation.

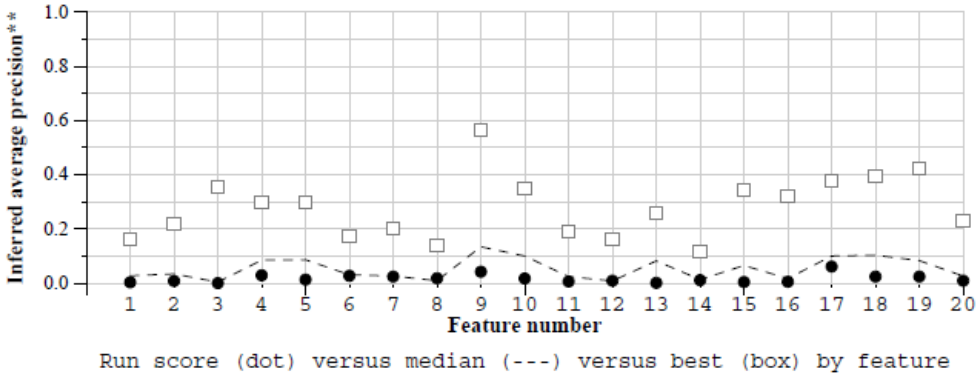

Figure 4. Inferred APs of pboost based method with 5,000 training shots.

5. CONCLUSIONS

In this paper, we described our HLF methods using the walk-based graph kernels and the pboost algorithm. For the TRECVID 2008 HLF task, the inferred APs of our method with the walk-based graph kernel were close to the median scores of the TRECVID 2009 HLF results. Our future work includes expansion of the pboost algorithm to reduce the computation cost and effective use of the image sequence information.

6. ACHKNOWLEDGEMENTS

The authors would like to thank Dr. Jean-Philippe Vert of Centre for Computational Biology, Mines ParisTech and Institut Curie for his close cooperation in the experiments using the walk-based graph kernels. Furthermore the authors wish to thank Dr. Koji Tsuda of AIST Computational Biology Research Center and Dr. Sebastian Nowozin of MPI for Biological Cybernetics for their valuable discussions and suggestions on the pboost algorithm and tool. This research was partially supported by Function and Induction Research Project, Transdisciplinary Research Integration Center, Research Organization of Information and Systems.

7. REFERENCES

[1] J.-P. Vert, T. Matsui, S. Satoh and Y. Uchiyama. High-level feature extraction using SVM with walk-based graph kernel. In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009).

[2] S. Nowozin, G. Bakir and K. Tsuda. Discriminative Subsequence Mining for Action Classification. In Proc. of Eleventh IEEE International Conference on Computer Vision (ICCV 2007).

[3] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. IEEE Trans. Pattern Anal. Mach. Intell., 23(8):800–810, Aug 2001.

[4] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. Int. J. Comput. Vision, 43(1):29–44, 2001.

[5] T. Gärtner. Exponential and Geometric Kernels for Graphs. In NIPS Workshop on Unreal Data: Principles of Modeling Nonvectorial Data, 2002.

[6] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized Kernels between Labeled Graphs. In T. Faucett and N. Mishra, editors, Proceedings of the Twentieth International Conference on Machine Learning, pp. 321–328. AAAI Press, 2003.

[7] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert. Graph kernels for molecular structure activity relationship analysis with support vector machines. J. Chem. Inf. Model., 45(4):939–51, 2005.

[8] Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), pp. 1–8. IEEE Computer Society, 2007.

[9] J. Ramon and T. Gärtner. Expressivity versus efficiency of graph kernels. In T. Washio and L. De Raedt, editors, Proc. of the First International Workshop on Mining Graphs, Trees and Sequences, pp. 65–74, 2003.

[10] P. Mahé and J.-P. Vert. Graph kernels based on tree patterns for molecules. Technical Report ccsd-00095488, HAL, September 2006.

[11] E. Aldea, J. Atif, and I. Bloch. Image classification using marginalized kernels for graphs. In Graph-Based Representations in Pattern Recognition, volume 4538/2007 of Lecture Notes in Computer Science, pp. 103–113. Springer Berlin/Heidelberg, 2007.

[12] J. Pei, J. Han, B. Mortazavi-Asl, J.Wang, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu. Mining Sequential Patterns by Pattern Growth: The Prefixspan Approach. IEEE Trans. on Knowl. and Data Eng., vol.16, no.10, pp.1424-1440, 2004.

[13] S. Nowozin. pboost sequence boosting toolbox. http://www.kyb.mpg.de/bs/people/nowozin/pboost/, 2007.