NTT Communication Science Laboratories and NII in TRECVID 2010

Instance Search Task

Takahito Kawanishi, Akisato Kimura, Kunio Kashino
NTT Communication Science Laboratories
NTT Corporation
Atsugi-shi, Kanagawa, Japan
{kawanisi, akisato, kunio} @eye.brl.ntt.co.jp

Shin'ichi Satoh, Duy-Dinh Le, Xiaomeng Wu, Sebastien Poullot
Multimedia Information Research Division
National Institute of Informatics
Chiyoda-ku, Tokyo, Japan
{satoh,ledduy,wxmeng,spoullot}@nii.ac.jp

Abstract—As a first step to the instance search task, we employed several existing image retrieval methods in combination (local feature matching, region matching, and global feature matching), according to the type and property of query images. The best run in this approach was ranked 23rd of 42 as regards the average precision result.

Keywords; PCA-SURF; Bag of Features; Color histogram; Haar-Wavelet; Face Detection; OpenCV

I. INTRODUCTION

The instance search task involves locating query topics from a collection of reference videos. The query topics consist of a set of about 5 example frame images, the regions containing the item of interest in the images, the video from which the images were selected, and an indication of the target type taken from the following set of strings: PERSON, CHARACTER, LOCATION, and OBJECT. One collection of reference videos consists of the Sound and Vision data from TRECVID 2009 and each video data is divided into many master shot references. The submitted data comprised 1000 candidates chosen from the master shots for each query topic. The score becomes high when the correct answer is put on the high rank.

The similar task with the instance search is the image retrieval from image database [1-3]. This time, we applied several basic existing image retrieval methods according to the type and property of query images. The types of query topic are "PERSON", "CHARACTER", "OBJECT", and "LOCATION". When the query topic type is "PERSON", the region of each query will include the face, and the face information will be important in terms of finding the topic. When the query type is "LOCATION", the region of the query may occupy almost all of the query image, and global information, for example color or frequency information will be useful. On the other hand, the query images are various sizes. Some query images have sufficient size but others have only a few features.

So, we adopted a method for selecting the most promising feature for the queries. When there are few features in the query region, we use a method for matching local features, when there are many features in the query region, we use a method for matching the bags of features as a region feature, and when the query region is as large as the query image, we use a method for matching global features.

First, we describe our features and their similarity measure in Section II. Section III provides an overview of our system. Section IV reports our submissions and results. Finally, we conclude by some remarks.

II. METHODS

We adopted methods that can be easily implemented with OpenCV library [4]. PCA-SURF features were used as the local features, a bag of PCA-SURF features was used when a reference image was matched with query topics, and a color histogram and a Haar wavelet feature were used as the global features. To select the object region from an image in the database, a face detection algorithm [5] is used in the OpenCV library. If there is no face region, we regarded the object region as entire image. The following sections describe the methods we used to generate and match the above features.

A. Local Feature Generation

PCA-SURF is used as a local feature that is similar to PCA-SIFT [6]. SURF [7] tested in this task has 256 dimensions. This dimension number is too large to identify deferent views of the same object on similar images. PCA is used to reduce the dimension number to 16. The similarity between the PCA-SURF features is defined as the normalized cosine similarity.

B. Matching from Local Region

A query object region is given in each query image. But if the query region includes the face, the face region is more informative than other regions. So, we use two query regions for each query image. If a face appears in the query region, the overlap between the query region and the face region is used to generate a region feature. If no face appears, the whole query region is used. On the other hand, no reference region in the reference images is given. If a face appears in the reference image, the face region is used as a reference region feature. If there is no face, the whole region in the reference image is used for the reference region feature.

If the query region has insufficient features, the PCA-SURF similarity is used as the similarity between query regions and reference images. We calculate all the

similarities for every combination between a point in the query region and a point in the reference region and we adopt the highest similarity as the region similarity between the query and the reference. If the query region has sufficient features, the BoF similarity is used as the region similarity. BoF is a histogram feature and a bin is a group of similar PCA-SURF features. The BoF similarity is defined as a histogram intersection. The number of bins is 1024. The histogram is not normalized because the reference region may be larger than the target object region when the region is equal to the entire reference image.

C. *Matching from Global Region*

Color and frequency features are used as global features. The color feature is a color histogram [8]. First, a color space is converted to HSV and each pixel in the region votes for the bin whose color is most similar to it. The bin number is 64.

The frequency feature is a Haar wavelet [9]. The Haar wavelet feature is a vector of 16 dimensions from a Haar wavelet image whose size is 4×4.

The similarity of the color histogram is an intersection and the histogram is not normalized. The similarity of the Haar wavelet feature is the normalized cosine similarity.

III. System Overview

This section describes the processing flow of the system tested in this task. The process is divided into three stages. In the first stage certain feature extraction parameters are learned. In the second stage, reference image features are generated. In the third stage, a query is given and a search algorithm is selected. After the search algorithm has been chosen, the similarities between the query and reference images are calculated and outputted. The following sections describe those stages in more detail.

A. *Learning Stage*

The learning stage generates the PCA-SURF parameters, the BoF (region) codebook, the BoF (face) codebook, and the color histogram codebook. This stage learns those parameters from over 3000 images and 1,000,000 SURF points contained in them. Each codebook is calculated using the k-means algorithm in OpenCV.

B. *Database Generation Stage*

First, 50 reference images are generated from each shot as each interval between the neighboring reference images is the same. Second, all of the features are generated from each reference image. The calculated features in each reference image are as follows.

a) PCA-SURF features

b) BoF (face)

c) BoF (whole image)

d) color histograms (global)

e) Haar wavelet (global)

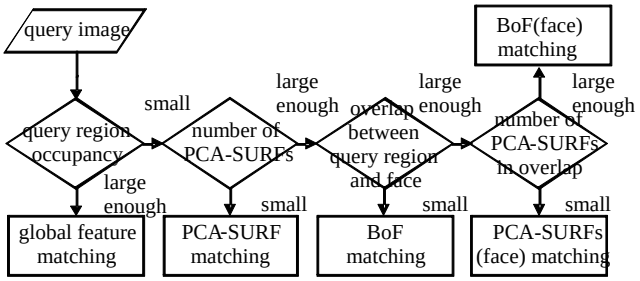


Figure 1. Feature Selection Algorithm in Search Stage.

C. *Search Stage*

1) *Feature Selection*

The feature selection algorithm in the search stage is shown in Figure 1. When a query image is given, a search algorithm and a feature for the query image are selected using the following procedure.

a) Does the query region occupy almost all of the query image?

Yes: use global features (whole query images vs. whole reference image)

No: go to b)

b) Are there sufficient PCA-SURF features?

Yes: go to c)

No: use PCA-SURF matching (query region vs. reference image)

c) Is the overlap between the query region and the face region large enough?

Yes: go to d)

No: use BoF matching (query region vs. whole reference image)

d) Are there sufficient PCA-SURF features in the overlaps?

Yes: use BoF matching (face region in query image vs. face region in reference image)

No: use PCA-SURF matching (face region in query image vs. face region in reference image)

2) *Search*

After the type of feature has been detemined for the query image, the query feature is generated. Then, the similarity between the query feature and the reference feature in each reference image is calculated. Finally the shot IDs are sorted by the similarity.

IV. TV2010 Submissions and Results

A. *Server Specifications*

TABLE I shows the specifications of the server used to process this task for about 180 hours of reference videos. For the learning stage, we use one process on one server. For the other stages, we use four servers and six processes on each server.

B. Processing Time

It took 850 minutes to generate features for all the reference images and it took 350 minutes to search the reference features for all the query images.

Since all the master reference features could not be held on memory. The features in a reference image were loaded on memory from a disk one by one. To reduce disk access time, one reference image was matched with all the query images rather than one query being matched with all the reference images. The search time for each query in the runs was obtained by dividing the searching time needed for all queries by the number of queries.

C. Runs

To make four runs, we tested two variations of thresholds in the feature selection procedure: "type-string sensitive" (Type) and "type-string unused" (Auto), and two variations of sorting key priority: "rank first" (Rank) and "score first" (Score). In the "type-string sensitive" condition, when the type was “PERSON” or “CHARACTER”, the threshold of procedure c) in Section 3 was set to lower than the default value so that BoF (face) and PCA-SURF descriptor (face) were more frequently selected and if the type was “LOCATION”, the threshold of procedure a) was set lower so that global features were more frequently selected. In the "type-string unused" condition, on the other hand, each threshold set to the default value and the feature was automatically selected regardless of the “type string” of the query topic.

In the "rank-first" sorting, the shot IDs in the run were sorted by the score first, whereas in the "score-first" sorting, they were sorted by the rank among each query image in the query topic.

TABLE I. SERVER SPECIFICATIONS.

CPU Type	Intel(R) Xeon(R) CPU X5460
CPU Frequency	3.16 GHz
CPU Cores	4
Memory	16G
OS	Cent OS 5.5

D. Results

In this task, we evaluated the average accuracy, elapsed time, and hits at several depths in the result by topics. Table II shows the result of our four runs and the best result in every submitted run across the topics. Our best result was ranked 23rd of 42 runs. If we select the best run for each team, the result of our NTT-NII team ranked 9th of 14 teams. Our four runs all produced similar scores.

Our rank of hits at a larger depth was better than that at a smaller depth. This is because our feature selection is relatively better than our ranking algorithm.

A comparison of our results and the best result by topic is shown in Figure 2. The best result does not realize a high score in every topic. But compared with the best result, our runs are not strong enough for every topic.

Red circles in Figure.2 indicate the topics where we obtained the best precision over the topics. Query images of topic 9020 and 9005 are shown in Figure. 3. Each top 4 outputs and those results for topic 9020 and 9005 are shown in Figure 4 and in Figure 5. PCA-SURF feature is used for topic 9020 in both “Type and Score” run and “Auto and Score” run. These runs could successfully find this topic in top 2 which consists of these small logos in Figure 3.

V. CONCLUDING REMARKS

We have described how we dealt with the instance search task this year. To make a baseline, we employed a set of basic existing methods in combination. However, the task was found to be very hard for most topics with the current strategy. We are now investigating the results in detail for future improvement of the system.

REFERENCES

[1] R. C. Veltkamp, M. Tanase and D. Sent, Features in content-based image retrieval systems: A survey, State-of-the-art in content-based image and video retrieval, pp. 97-124 1999.

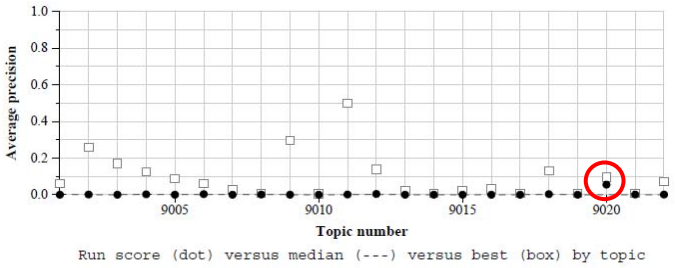
[2] R. Schettini, G. Ciocca and S. Zuffi, A survey on methods for colour image indexing and retrieval in image databases. In: R. Luo and L. MacDonald, Editors, Color Imaging Science: Exploiting Digital Media, Wiley, New York 2001.

[3] S. Antani, R. Kasturi and R. Jain, A survey on the use of pattern recognition methods for abstraction, indexing, and retrieval of images and video, Pattern Recognit., vol. 35, pp. 945 2002.

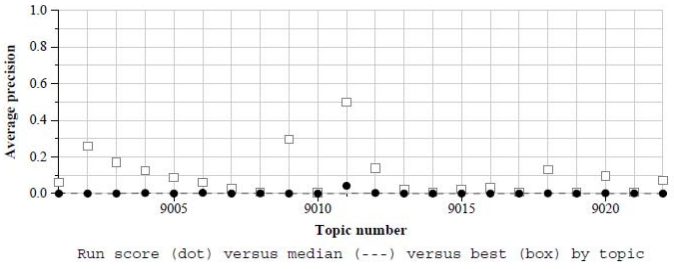
[4] http://www.sourceforge.net/projects/opencvlibrary/.

TABLE II. RUN RESULT VS. BEST SCORED RESULT

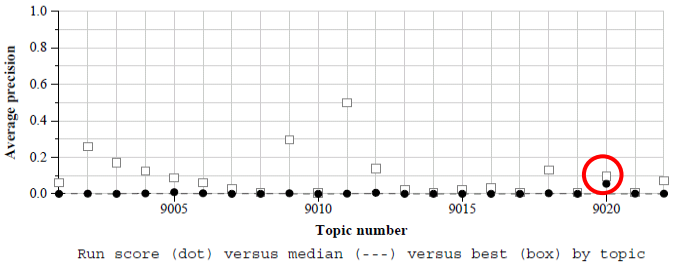
	Average precision (rank)	Elapsed time [mins] (rank)	Hits at depth X in the result set (rank)			
			10	30	100	1000
Type + Score	0.327%(26th)	16(28th)	0.136(26th)	0.182(32nd)	0.636(30th)	3.32(21st)
Type + Rank	0.259%(28th)	16(28th)	0.0909(32nd)	0.136(34th)	0.636(30th)	3.27(22nd)
Auto + Score	0.386%(25th)	16(28th)	0.182(24th)	0.273(27th)	0.773(25th)	4.82(14th)
Auto + Rank	0.445%(23rd)	16(28th)	0.136(26th)	0.273(27th)	0.864(19th)	4.27(19th)
Best	53.436%	0.000	8.409	18.227	33.773	35.318



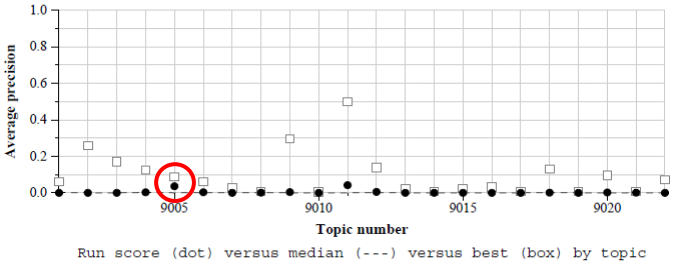
(a) Type and Score



(b) Type and Rank



(c) Auto and Score



(d) Auto and Score

Figure 2. Our runs vs. best by topics.

[5] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, Proc. IEEE Computer Vision and Pattern Recognition, 2001, p. 511.

[6] Y. Ke and R. Sukthankar, PCA-SIFT: A more distinctive representation for local image descriptors" , Proc. Conf. Computer Vision and Pattern Recognition, pp. 511-517, 2004.

[7] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF). Comp. Vision and Image Understanding 110(3), 346–359 (2008)

[8] M.J. Swain and D.H. Ballard, Color indexing. Intl. J. Comput. Vis. 7 1 (1991), pp. 11–32.

[9] C. Brambilla, A. Della Ventura, I. Gagliardi, R. Schetini, Multiresolution wavelet transform and supervised learning for content-based image retrieval, IEEE International Conference on Multimedia Computing Systems, Vol. 1, 1999, pp. 183–188.

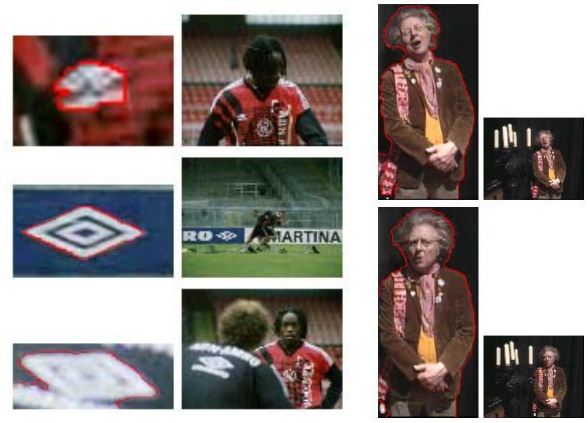


Figure 3. Query images of topic 9020 (left) and topic 9005 (right).

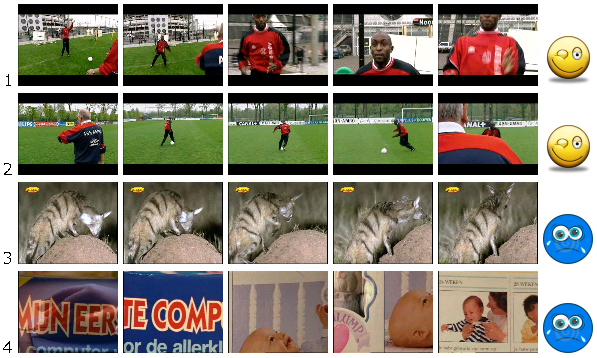


Figure 4. Top 4 outputs and those results of topic 9020 in our "Type and Score" run.

Figure 5. Top 4 outputs and those results of topic 9005 in our "Auto and Score" run.