NTT Communication Science Laboratories and NII at TRECVID 2011

Instance Search Task

Takahito Kawanishi, Kunio Kashino
NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato-Wakamiya Atsugi-shi, Kanagawa, Japan
{kawanishi.takahito, kashino.kunio}@lab.ntt.co.jp

Yong Qing Sun
NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikari-no-Oka, Yokosuka-shi, Kanagawa, Japan
yongqing.sun@lab.ntt.co.jp

Shin'ichi Satoh, Duy-Dinh Le, Caizhi Zhu
Multimedia Information Research Division, National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
{satoh,ledduy,cai-zhizhu}@nii.ac.jp

***Abstract*—We tried using several local and global features for the TRECVID 2011 instance search task. From last year's experience, we understood that face features were effective for this task. So we tried using a well-known face recognition algorithm for this task, and obtained better results for certain topics. Unfortunately our runs included some bugs, so the results were worse than last year's. Our best run ranked 25th in 37 runs as regards the average precision result.**

Keywords; color histogram; local feature; face recognition

I. INTRODUCTION

An instance search task involves locating query topics from a collection of reference videos. The query topics consist of a set of about 2-6 query source images, target images that are the regions containing the item of interest in the query source images, and an indication of the target type taken from the following set of strings: PERSON, LOCATION, and OBJECT. One collection of reference videos consists of BBC rushes amounting to 20982 movie shots including some very similar retakes. The submitted data comprised 1000 candidates chosen from the reference videos for each query topic. The score becomes high when the correct answer ranks high.

A similar instance search task involves image retrieval from an image database [1-3]. Here, we employed several basic existing image retrieval methods according to the characteristics of the query images in each query topic.

II. METHODS

We adopted methods that can be easily implemented with OpenCV library [4]. Face, local and global descriptors were used for matching query images to reference key frames. If there was a face region in the query target images, the feature suitable for the face was selected. If the query target image have rich features (over 50 SURFs), the feature in the query target image was used. If few features were found in the query target image, a global feature and a query source image had to be used.

A. Face descriptors

We selected 10 key frames per reference video and extracted the frontal face region from query target images and reference key frames using a Viola-Jones face detector [5] in OpenCV. The facial descriptor [6] was formed by extracting the pixel intensity around facial points. 9 facial feature points were detected, and 4 more facial feature points were inferred from these 9 points. In total, there were 13 feature points from which features were extracted. The features were intensity values lying within the circle with radius of 15 pixels. The output feature had 13x149 = 1,937 dimensions. We computed the similarity between faces extracted from target images provided by the query and faces extracted from key frames of the reference videos and then ranked the shots using the similarity scores. To avoid the effect of glasses or masks, we can also select the 3 most similar facial features whose dimensions are 3x149 = 447.

B. Local descriptors

SURF [7] was used as a local descriptor. In advance, we selected 10 key frames from reference videos, extracted reference SURFs from each key frame and stored the reference SURFs in the database. When a query image was given, query SURFs were extracted from the query target images. The reference SURF in the database nearest to each query SURF in a query image was selected. Similar reference videos were decided based on the distances between the nearest reference SURFs and the query SURFs. One measure between the reference video and the query topic is the number of nearest SURFs in each reference video. Another measure is the shortest distance between the query SURFs and the reference SURFs in each reference video.

C. Global descriptors

In addition to the above two descriptors, color and the frequency features were used as global features. A color feature is a color histogram [8]. A frequency feature is a histogram of SURFs (a bag of SURFs). The bin of each histogram is generated by vector quantization algorithm. To learn a VQ centroid, we used about 10000 images from the Sound and Vision 2009 dataset.

The similarity between these histograms is the intersection between the histogram of the query images and that of the reference key frames. The similarity measure between a reference video and a query topic is the greatest

similarity between the reference key frames and the query images.

III. TV2011 SUBMISSION

A. *Submitted runs*

We use different methods when we find faces, when we find over 50 SURFs per an image, and for any other case. Table 1 shows our method for each status. “All facial parts” means that we use a face descriptor of all 13 facial points. “3 facial parts” means that we use that of the 3 nearest facial points. “Color” means a color histogram feature. “BoF” means a SURF histogram. “LD” means a local descriptor whose distance measure is “NUM” or “MIN”. “NUM” means the number of the nearest LD and is used as a similarity measure and “MIN” means the shortest distance between the query LDs and the reference LDs in each reference key frame and is used as a distance measure. “Target” means that we used a query target image and “source” means that we used a query source image.

In this experiment, we used four IBM SYSTEM3550M2 servers which each had two CPUs and a 24Gbyte memory. The CPU was a Xeon X5570 (Quad-Core, Hyper-Threading, 2.93GHz). To learn VQ codes, we used one process on one server. For other cases, we used four servers and 16 processes on each server.

B. *Results and bugs*

In this task, we evaluated the average accuracy and elapsed time. Table I shows the average result for every submitted run across the topics. Our best run was ranked 25th of 37 runs. RUN2 had a critical bug and was not evaluated. RUN1, 3, 4 also had a serious bug that eliminated the id of the shot with the top score so the average precision became much worse.

Figure 1 compares our results and the best result by topic. The red circles in Fig. 1 indicate the topics for which we obtained the best precision. The top 4 outputs and the results for topic 9039 are shown in Fig. 2. The run used all the facial parts and could successfully find this topic in top 2 which consists of a frontal face.

IV. CONCLUDING REMARKS

We have described how we dealt with the instance search task this year. To establish 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 with a view to improving the system.

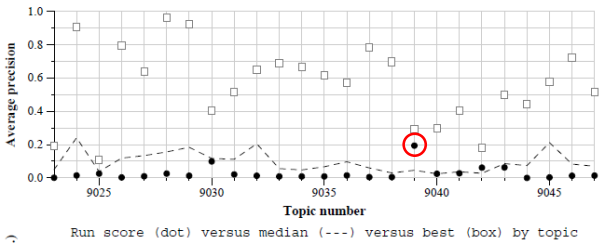

Figure 1. Our best runs vs. best by topics.

Figure 2. Our 4 outputs in our best topic (red circle in Figure 1)

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/.

[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] M. Everingham, J. Sivic, and A. Zisserman, “’Hello, My name is... Buffy’- automatic naming of charecters in tv video,” in Proc. British Machine Vison Conf., 2006.

[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.

TABLE I. RUN SPECIFICATION AND RESULTS

	Face found	Over 50 SURFs	use global feature	Average Precision(rank)	Elapsed time[min.] (rank)
RUN1	All facial parts	Color(target)	Color(source)	0.02724(25th)	0.208(6th)
RUN2	3 facial parts	LD(NUM, target)	BoF(source)	DNF	DNF
RUN3	LD(MIN, target)	LD(MIN, target)	LD(MIN,source)	0.00476(36th)	4.288(16th)
RUN4	Color(target)	BoF(target)	LD(NUM,source)	0.00868(34th)	2.1(12th)