NTT Communication Science Laboratories at TRECVID 2014 Instance Search Task

Masaya Murata, Hidehisa Nagano, Kaoru Hiramatsu,
Takahito Kawanishi, Kunio Kashino
NTT Communication Science Laboratories (NTT CSL), NTT Corporation
3-1, Morinosato Wakamiya, Atsugi-Shi, Kanagawa, 243-0198, Japan
email: murata.masaya@lab.ntt.co.jp

Abstract

This paper reports our method and experimental result on the TRECVID 2014[1] instance search task. Since 2012, we have been applying BM25 (Best Match 25), i.e., the state-of-the-art probabilistic information retrieval method in the field of text retrieval, to the instance search task. The standard BM25 uses the well-known Inverse Document Frequency (IDF) as the key-point discriminative power, and in 2012 we have employed this methodology for the instance search task. The evaluation result was promising and we continued this research direction. In 2013, we proposed an approach to enhance the discriminative power of IDF and devised a new weight called exponential IDF (EIDF). The BM25 with EIDF method is called exponential BM25 (EBM25) and together with the region-of-interest (ROI) effect, our approach recorded the second-ranked search accuracy among all of the submission runs. This year, we also proposed a video re-ranking method using ROI images and executed it for video search results obtained by the EBM25 with ROI method. Our result this year was ranked third among all of the submission teams, indicating the effectiveness of the proposed approach.

I. Instance Search using EBM25 with ROI

We first review our video retrieval method proposed to address the instance search task last year[2]. The video ranking function is summarized as follows:

$$p(\mathrm{REL}=\mathrm{rel}|v,q) \propto_q \sum_{q_i>0} \frac{v_i}{v_i + 2(0.25 + 0.75(vl/avvl))} w_i r_i, \tag{1}$$

$$\text{where } w_i = \log\left(\frac{e^{-n_i/\lambda}(N - n_i + e^{n_i/\lambda} - e^{-n_i/\lambda} + 1)}{(e^{n_i/\lambda} - e^{-n_i/\lambda} + 1)(n_i + e^{-n_i/\lambda})}\right), \tag{2}$$

$$r_i = \begin{cases} 10 & (\text{if query feature } i \text{ appears in ROI}) \\ 1 & (\text{otherwise}) \end{cases} \tag{3}$$

Here, REL is the probabilistic event taking binary events of rel or iirel which means relevance of the video v to the query image q or irrelevance of the video v to the query image q, respectively. Therefore, the left side of eq. (1) denotes the conditional relevance probability of v to q. $\propto_q$ is the ranking equivalence sign with respect to q which implies that the video ranking result in the decreasing order of the left side of the equation becomes the same as that in the decreasing order of the right side of the equation. v and q are n-dimensional random vectors whose elements are within frequencies (within video and within query image frequencies) of features indexed from 1 to n, where n is the total feature number under consideration. We can also mention that the n is the size of the vocabulary of interest. For example, when dealing with content-based video retrieval tasks, local image features such as SIFT features are often taken and their indices are defined by the representative visual words in the pre-defined codebook. Then, the aforementioned vocabulary corresponds to the visual words in the codebook. The q_i in the right side of eq. (1) denotes the ith vocabulary frequency within q and $\sum_{q_i>0}$ indicates the summation over vocabularies whose frequencies within q are over 0. We call these vocabularies query features. In the same way, v_i denotes the ith vocabulary frequency within v. vl is the video length defined by $\sum_{q_i \geq 0} v_i$ and $avvl$ is the average vl over all of the stored videos in the database. Note that the video length is defined by all of the vocabularies under consideration, not only by the query features $q_i > 0,\ i = 1, 2, \cdots, n$.

w_i in the right side of eq. (1) is called the discriminative power of the query feature i. Its standard choice is the well-known IDF, however, we rather use the exponential IDF (EIDF) expressed in eq. (2). The reason is that since the IDF statistically estimates the discriminative

power, it requires a sufficient amount of data to make the estimated power meaningful. Especially in the content-based video retrieval tasks, local image features such as SIFT features are described by high-dimensional vectors and the codebook size (total number of visual words) is set sufficiently large to reduce the vector quantization errors. Because such assumption on the sufficiently large data does not generally hold, we suppose a different assumption that the query feature is less discriminative when it shows a frequently-appearing tendency in a relatively small data, leading to the derivation of EIDF. In the TRECVID 2013 instance search task, we experimentally showed that the BM25 with EIDF approach dramatically increased the instance search accuracy.

The r_i is the ROI effect and it takes the binary values of 10 when the query feature i appears in the ROI or otherwise takes 1 as shown in eq. (3). Without this ROI factor, our instance search results become similar to the search results obtained by an conventional similar video retrieval method. We can also define the new weight by regarding $w_i r_i$ as wr_i, that is, by regarding the two factors as one new factor. In the next section, we explain the different formulatios for the wr_i.

II. Discriminative Power Formulation

When $r_i = 1$ for all of the query frequencies, that is, when the ROI effect is not taken into accout, then $wr_i = w_i$ and it is expressed as follows[3]:

$$w_i = \log\left(\frac{p(e_i|\mathrm{rel})p(\bar{e}_i|\mathrm{irrel})}{p(e_i|\mathrm{irrel})p(\bar{e}_i|\mathrm{rel})}\right) \tag{4}$$

$$\approx \log\left(\frac{E[\xi|R, r_i](1 - E[\xi|N - R, n_i - r_i])}{E[\xi|N - R, n_i - r_i](1 - E[\xi|R, r_i])}\right) \tag{5}$$

Here, the two conditional probabilities $p(e_i|\mathrm{rel})$ and $p(\bar{e}_i|\mathrm{irrel})$ denote $p(\mathrm{ELITE}_i{=}\ \mathrm{elite}|\mathrm{REL} = \mathrm{rel})$ and $p(\mathrm{ELITE}_i{=}\ \mathrm{non\ elite}|\mathrm{REL} = \mathrm{irrel})$, respectively. ELITE_i is a random event for the query feature i taking binary events of either elite or $\mathrm{non\ elite}$. This eliteness attribute is, in other words, explained by the aboutness or star property. Therefore, when an certain feature becomes star, this feature is expected to appear frequently in videos. $p(e_i|\mathrm{rel})$ can be interpreted as the probability of query feature i becoming elite given the relevant video set since $p(\mathrm{rel}) = \sum_{(v,q)} p(\mathrm{rel}|v, q)p(v, q)$, where the summation is over all of the video and query set. On the other hands, $p(\bar{e}_i|\mathrm{irrel})$ can

be interpreted as the probability of query feature i becoming non elite given the iirrelevant video set. Note that $p(e_i|\mathrm{rel}) + p(\bar{e}_i|\mathrm{rel}) = 1$ and $p(e_i|\mathrm{irrel}) + p(\bar{e}_i|\mathrm{irrel}) = 1$ hold, respectively.

Then we suppose that $p(e_i|\mathrm{rel}) \approx E[\xi|R, r_i]$, where $\xi \equiv p(e_i)$, and R and r_i denote the number of relevance videos and the number of relevant videos containing the query feature i. Therefore, $E[\xi|R, r_i]$ is the expectation for the probability of query feature i becoming elite given R and r_i. In the same way, we suppose that $p(e_i|\mathrm{irrel}) \approx E[\xi|N - R, n_i - r_i]$, where N and n_i denote the number of videos and the number of videos containing the query feature i, respectively. These two expectations are obviously different and they can be calculated by the following Bayes' rules.

$$p(\xi|R, r_i) = \frac{p(r_i|\xi, R)p(\xi|R)}{\int p(r_i|\xi, R)p(\xi|R)d\xi} \tag{6}$$

$$p(\xi|N - R, n_i - r_i) = \frac{p(n_i - r_i|\xi, N - R)p(\xi|N - R)}{\int p(n_i - r_i|\xi, N - R)p(\xi|N - R)d\xi} \tag{7}$$

Here, we can interpret that $p(\xi|R) = \sum_{r_i=0}^{R} p(\xi|R, r_i)p(r_i)$ and $p(\xi|N-R) = \sum_{(n_i-r_i)=0}^{N-R} p(\xi|N-R, n_i - r_i)p(n_i - r_i)$. Therefore, if $p(\xi = r_i/R|R, r_i) = 1$ and $p(r_i = 0) = p(r_i = 1) = \cdots = p(r_i = R) = 1/(R + 1)$, $p(\xi|R) \sim 1/(R + 1)$ which indicates that $p(\xi|R)$ follows a discrete uniform distribution with probability of $1/(R+1)$. In the same way, $p(\xi|N-R) \sim 1/(N-R+1)$, indicating that $p(\xi|N - R)$ follows a discrete uniform distribution with probability of $1/(N - R+1)$. We can extend these ideas and suppose that the two probabilities $p(\xi|R)$ and $p(\xi|N-R)$ follow continuous uniform distributions, which is special cases of Beta distributions denoted as $\mathrm{Beta}(1, 1)$. This setting leads to the derivation of the well-known IDF. We can further extend this idea and suppose that these two probabilities follow arbitrary Beta distributions denoted as $\mathrm{Beta}(\alpha, \beta)$ and $\mathrm{Beta}(\alpha', \beta')$, respectively. Together with the assumptions that $p(r_i|\xi, R) \sim \mathrm{Binomial}(R, \xi)$, $p(n_i - r_i|\xi, N - R) \sim \mathrm{Binomial}(N - R, \xi)$, the left sides of eqs. (6) and (7) become $p(\xi|R, r_i) \sim \mathrm{Beta}(r_i + \alpha, R - r_i + \beta)$ and $p(\xi|N - R, n_i - r_i) \sim \mathrm{Beta}(n_i - r_i + \alpha', N - R - (n_i - r_i) + \beta')$, respectively. Therefore, these two expectations become

$$E[\xi|R, r_i] = \frac{r_i + \alpha}{R + \alpha + \beta} \tag{8}$$

$$E[\xi|N - R, n_i - r_i] = \frac{n_i - r_i + \alpha'}{N - R + \alpha' + \beta'} \tag{9}$$

Here, note that the expectation for a random variable following $\mathrm{Beta}(\alpha, \beta)$ is $\alpha/(\alpha + \beta)$. Then,

the resulting w_i is expressed as follows:

$$w_i \approx \log\left(\frac{(r_i+\alpha)(N-R+\beta'-n_i+r_i)}{(n_i-r_i+\alpha')(R+\beta-r_i)}\right) \tag{10}$$

For eq. (10), setting $R=r_i=0$, and $\alpha=\alpha'=e^{-n_i/\lambda}$ and $\beta=\beta'=e^{n_i/\lambda}-e^{-n_i/\lambda}+1$ yields the following EIDF:

$$w_i^{\mathrm{EIDF}} = \log\left(\frac{e^{-n_i/\lambda}(N-n_i+e^{n_i/\lambda}-e^{-n_i/\lambda}+1)}{(e^{n_i/\lambda}-e^{-n_i/\lambda}+1)(n_i+e^{-n_i/\lambda})}\right) \tag{11}$$

Here, the λ is a design parameter which depends on N, however, we often take $\lambda = 100 \sim 1000$. As the n_i becomes large, the w_i rapidly becomes small. Therefore, contributions from query features showing frequently-appearing tendency to the video ranking scores are sufficiently suppressed by using EIDF as the discriminative powers.

III. Ranking based on Joint Relevance Probability

This year, we also considered $p(\mathrm{REL^m}=\mathrm{rel}|v,m)$, where $\mathrm{REL^m}=\mathrm{rel}$ is the relevance of v to m and m is the random vector whose elements are within ROI query feature frequencies. Therefore, compared to q, m only describes query features appearing within the ROI and we call such query features as ROI features. Then the following equation holds.

$$p(\mathrm{REL}=\mathrm{rel}|v,q)p(\mathrm{REL^m}=\mathrm{rel}|v,m) = p(\mathrm{REL}=\mathrm{rel},\mathrm{REL^m}=\mathrm{rel}|v,q,m) \tag{12}$$

Here, the right side of eq. (12) denotes the joint relevance probability given v, q and m. Then, we can state that it might be better to rank videos according to the right side of eq. (12) than to rank videos in the decreasing order of $p(\mathrm{REL}=\mathrm{rel}|v,q)$ in eq. (1). However, the problem is that we can not use the right side of eq. (1) with q replaced by m for $p(\mathrm{REL^m}=\mathrm{rel}|v,m)$, since eq. (1) is not an equality. Therefore, we define $p(\mathrm{REL^m}=\mathrm{rel}|v,m)$ as follows:

$$\begin{aligned}
\mathrm{BM25}(v,q) &\equiv \sum_{q_i>0}\frac{v_i}{v_i+2(0.25+0.75(vl/avvl))}w_i r_i \\
\mathrm{BM25}(v,m) &\equiv \sum_{m_j>0}\frac{v_j}{v_j+2(0.25+0.75(vl/avvl))}w_j \\
p(\mathrm{REL^m}=\mathrm{rel}|v,m) &= \frac{\mathrm{BM25}(v,m)}{\max_v \mathrm{BM25}(v,m)}
\end{aligned} \tag{13}$$

Here, $\max_v \mathrm{BM25}(v,m)$ is the maximum $\mathrm{BM25}(v,m)$ among all of the videos in the database with respect to the ROI features m. In other words, it is the ranking score for the 1st ranked

video retrieved by using m. It is therefore always 1 or less and $p(\mathrm{REL}^{\mathrm{m}} = \mathrm{rel}|v, m)$ for the first ranked video is always 1. However, since the eq. (13) becomes quite small as the ranking score for the 1st ranked video becomes large, we define its lowest value as θ. We therefore set $p(\mathrm{REL}^{\mathrm{m}} = \mathrm{rel}|v, m)$ under θ as θ. We set $\theta = 0.3$ for this year's instance search task.

The video ranking score based on this joint relevance probability is expressed as follows:

$$p(\mathrm{REL} = \mathrm{rel}|v, q)p(\mathrm{REL}^{\mathrm{m}} = \mathrm{rel}|v, m) \propto_{(q,m)} \begin{cases} \frac{\mathrm{BM25}(v,q)\mathrm{BM25}(v,m)}{\max_v \mathrm{BM25}(v,m)} & (\text{if eq. } (13) \ \geq \ \theta) \\ \theta \times \mathrm{BM25}(v, q) & (\text{otherwise}) \end{cases} \tag{14}$$

We can also mention that eq. (14) is a re-ranking method of video search results using ROI features. This re-ranking method is not necessarily performed for all of the original video search results obtained by using $\mathrm{BM25}(v, q)$. Indeed, this year we performed re-ranking of the top 20 video search results to boost the instance search accuracy further.

IV. Instance Search System

Our instance search system is composed of five procedures to generate the final search results for each instance topic. We explain each part in the following sections.

1. Query feature extraction

We extracted key-points from instance topic query images using the Harris-Laplace detector[4], and featured the key-points on the basis of the 128-dimensional SIFT[5] and 192-dimensional compact CSIFT methods[6][7]. We thus used two local features: luminance and chrominance. These query features constitute our visual vocabulary. Duplicate query features were later removed from the vocabulary.

2. Shot feature extraction and feature matching

The key-frame images were extracted, at the rate of one frame per second, from about $470,000$ videos shots in the database (DB), and we then obtained about a total of $4,100,000,000$ (four billion and one-hundred million) visual key-points featured with both the SIFT and CSIFT vectors. These key-points (called shot features) were matched against the query features according to the cosine value between the two feature vectors. Then the feature pair showing the highest cosine value larger than 0.9 was considered as matched. These feature matched results were then used to count the query feature frequencies within each video shot.

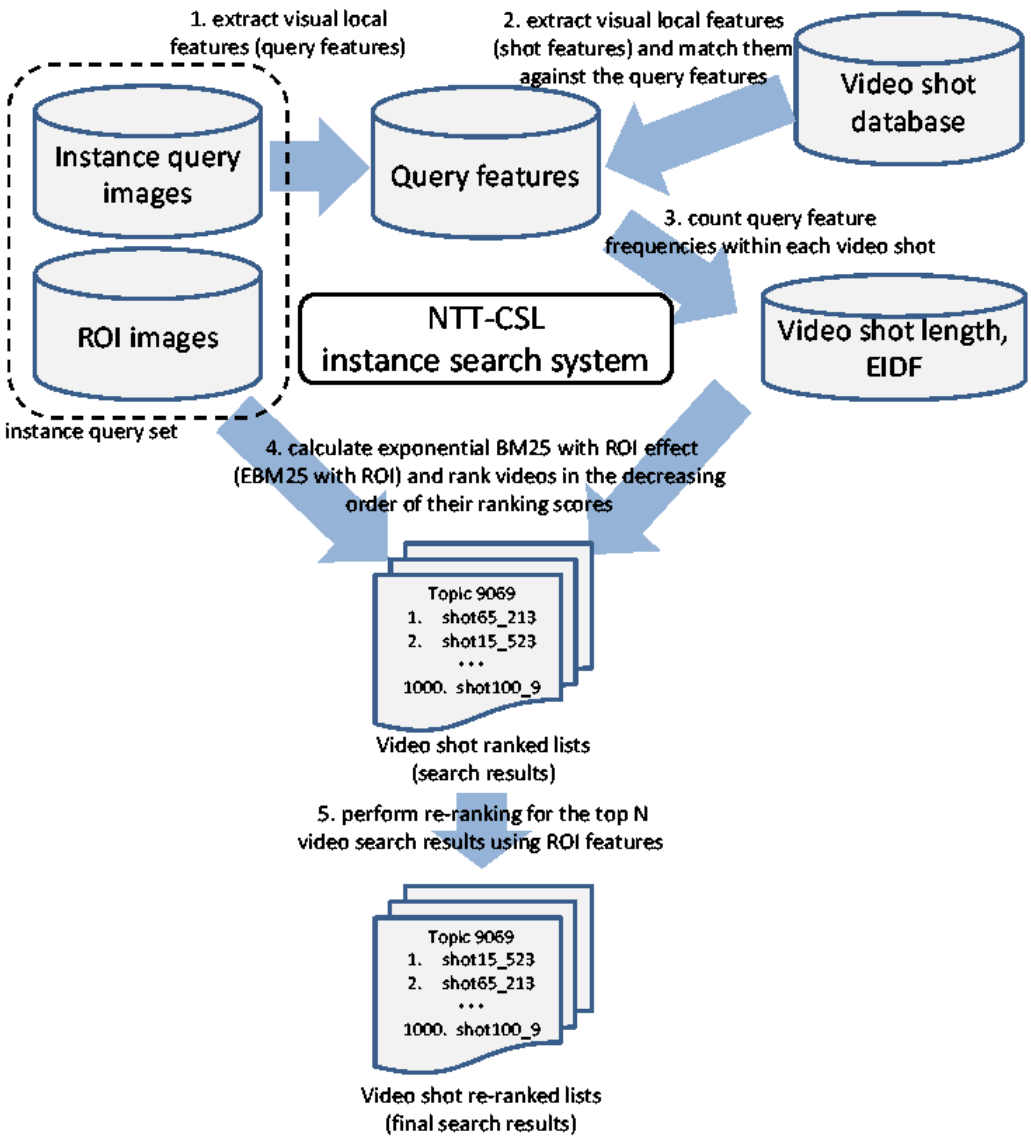

Fig. 1: NTT-CSL instance search system.

3. Video shot length and EIDF

From the feature matching results, we obtained the query feature frequencies (occurrence numbers) within each video shot. As already explained in Section I, we use v_i to denote the frequency of query feature i within each video shot. Note that since we used two features such as SIFT and CSIFT, video shot length vl and the average number $avvl$ were calculated twice based on these two features by the method explained in Section I. EIDFs for all of the query

features were also calculated for both SIFT and CSIFT descriptors by the method explained in Section II. The counted and calculated values v_i and w_i were used afterward to calculate the shot ranking scores of the EBM25 with ROI.

4. Instance search using EBM25 with ROI

After identifying the query features appearing within the ROI images (called ROI features), the shot ranking scores of the EBM25 with ROI (eq. (1)) were calculated for all of the video shots in the database. Again, these scores are calculated for both SIFT and CSIFT descriptors and they are later added together to generate the video shot ranked lists (search results).

5. Search results re-ranking

We finally performed the video re-ranking (eq. (14)) for the top 20 search results obtained by the EBM25 with ROI method. Since this re-ranking procedure was designed to raise videos containing many ROI features to higher ranks, it is expected to improve the instance search accuracy of the original search results. In the next section, we show the evaluation result for our submittied run this year.

V. SUBMITTED RUN AND EVALUATION RESULTS

We only submitted one run this year where the parameters in eq. (14) were set as $\lambda = 100$ and $\theta = 0.3$. Figure 2 shows the overall evaluation results for 22 submission teams this year. The

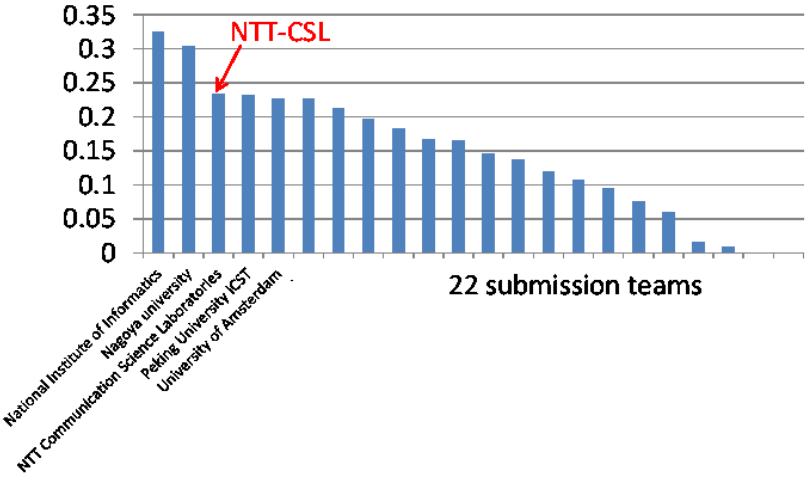

Fig. 2: Instance search accuracy measured in terms of MAP.

vertical axis is the mean average precision (MAP) measuring the instance search accuracy and the horizontal axis is the submission teams. Note that the maximum MAP among each team's submitted runs is shown in the figure. National Institute of Informatics (NII) scored the highest MAP and Nagoya University scored the second-highest MAP. We, NTT Communication Science Laboratories (NTT CSL), were ranked 3rd, and Peking University and University of Amsterdam followed. Although the MAPs of the top 2 teams were significantly higher than ours, we confirmed the effectiveness of our instance search approach to some extent. Further comparison results such as between our method and that without the proposed re-ranking procedure would be provided at the workshop held at the beginning of November this year. Figure 3 shows the detailed evaluation results for our submitted run provided by the NIST.

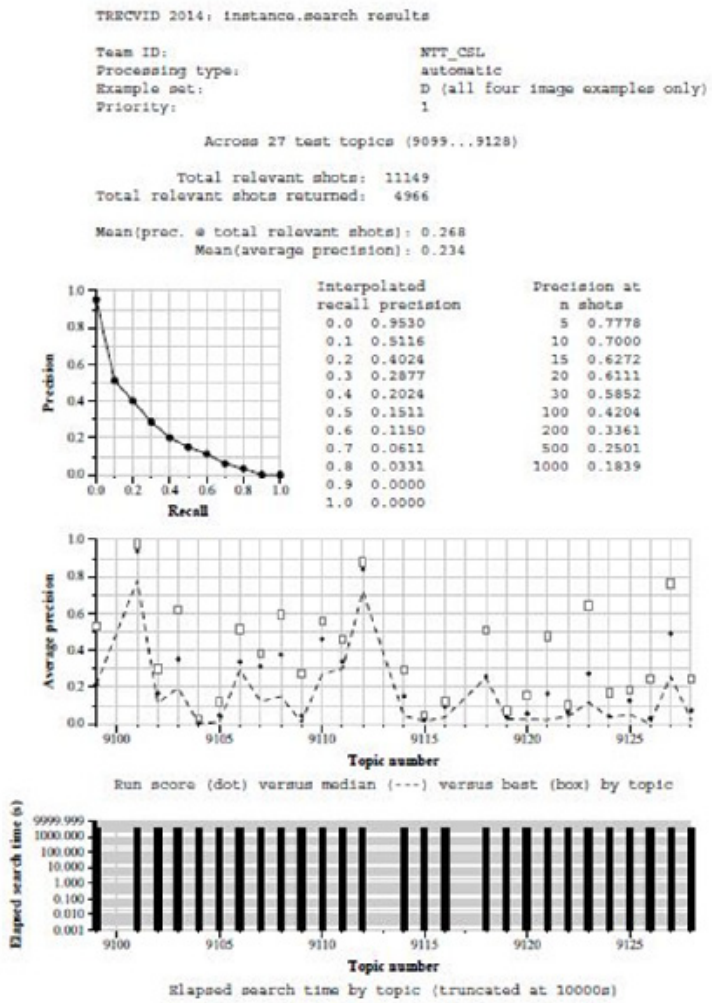

Fig. 3: Detailed result of NTT-CSL run.

VI. Conclusion

This year we have proposed a new video re-ranking method based on the joint relevance probability and executed it using ROI features in order to raise videos containing instance topics to higher ranks in the search result rankings. The evaluation results have confirmed its performance to some extent, however, more effective use of the ROI features is now under exploration. We should finally mention that for instances from which characteristic query features are not sufficiently extracted, our search methodology does not seem to work well. Therefore, as well as the way of using ROI images, an approach that is capable of searching for such difficult instances is also highly desired.

References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, G. Queenot, "TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics", Proc. of TRECVID 2014, 2014.

[2] M. Murata, T. Izumitani, H. Nagano, K. Kashino, S. Satoh, "NTT Communication Science Laboratories and National Institute of Informatics at TRECVID2013 Instance Search Task", Proc. of TRECVID 2013, 2013.

[3] M. Murata, H. Nagano, R. Mukai, K. Kashino, S. Satoh, "BM25 with Exponential IDF for Instance Search", IEEE Trans. on Multimedia, 2014.

[4] Harris, C., Stephens, M., "A combined corner and edge detector.," *4th Alvey Vision Conference*, pp. 147–151, 1988.

[5] Lowe, D., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, 60(2), 91–110, 2004.

[6] Mikolajczyk, K., Schmid, C., "Scale and affine invariant interest point detectors.," *International Journal of Computer Vision*, 60(1), 63–86, 2004.

[7] Zhu, C. Z., Sato, S., "Large Vocabulary Quantization for Searching Instances from Videos.," *Proc. of ICMR'12*, No. 52, 2012.