NTT Communication Science Laboratories and National Institute of Informatics at TRECVID 2012 Instance Search and Multimedia Event Detection Tasks

Masaya Murata†, Tomonori Izumitani†, Hidehisa Nagano†, Ryo Mukai†,
Kunio Kashino†, and Shin'ichi Satoh‡
† NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato Wakamiya, Atsugi-Shi, Kanagawa, 243-0198, Japan
‡ National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, 101-8430, Japan
email: {murata.masaya, izumitani.tomonori}@lab.ntt.co.jp

October 29, 2012

ABSTRACT

This paper reports our methods and experimental results for two TRECVID 2012 tasks of instance search and multimedia event detection. For the instance search, we applied the BM25(Best Match 25) method, the state-of-the-art ranking function in the field of text retrieval, to the video retrieval task. Our BM25 features the following three factors: (i) the keypoint importance degree (ii) keypoint frequency in the video, and (iii) normalization by the video length. The keypoint importance degrees are the weights of the local features detected from instance topic images. They were estimated by two different approaches: using the instance search video collection and using google images search results obtained by inputting the topic keywords. We defined the four submission runs for this task as follows: (a) NTT-NII_1 run - applying BM25 with IFF(Inverse Frame Frequency) weight; (b) NTT-NII_2 run - applying BM25 with RSJ(Robertson-Sparck-Jones) weight; (c) NTT-NII_3 run - the same as above, but slightly different parameter values; and (d) NTT-NII_4 run - NTT-NII_1 run combined with a NII's independent run. Here, IFF and RSJ are the methods of estimating the weights on keypoint importance degrees. Note that the details of NTT-NII_4 run will be omitted in this paper. The overall experimental results showed that NTT-NII_1 scored the

highest search accuracy among our four submission runs and that the use of BM25 as the baseline ranking approach for the instance search task is promising. For multimedia event detection, our system utilized text data in addition to image data. For image features, we tried SIFT-based local features and color histograms separately. In the training phase, a space was determined by integrating image and text information using canonical correlation analysis (CCA). Video frames were transformed into the obtained space and an SVM-based classifier was applied for each event in the detection phase. The evaluation results indicate that CCA-based image and text information integration does not work effectively when only the provided text meta data are used.

1. INSTANCE SEARCH: NTT-NII_1 RUN

1.1 System Overview

Figure 1 shows an overview of our NTT-NII_1 instance search system. To improve the search accu-

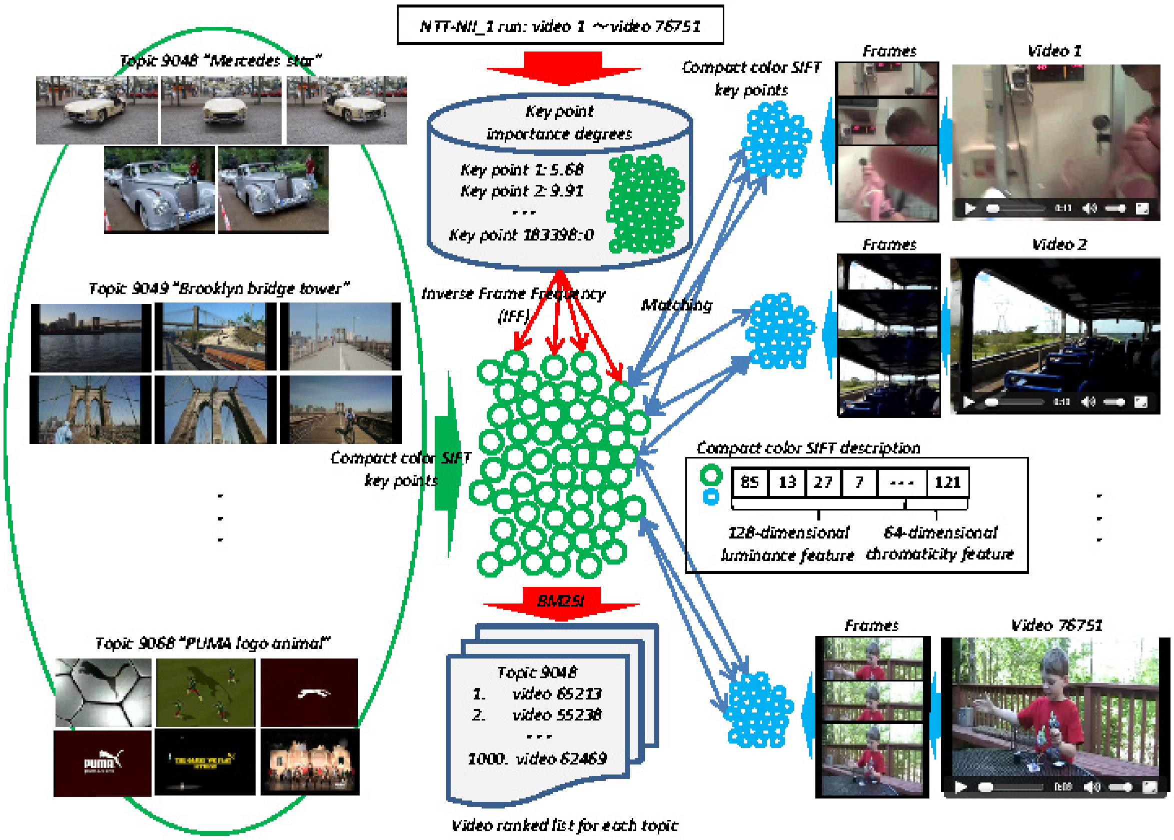

NTT-NII_1 run: video 1 ~video 76751
Topic 9048 "Mercedes star"
Key point importance degrees
Key point 1: 5.68
Key point 2: 9.91
...
Key point 183398:0
Compact color SIFT key points
Frames
Video 1
Topic 9049 "Brooklyn bridge tower"
Inverse Frame Frequency (IFF)
Matching
Frames
Video 2
Compact color SIFT key points
Compact color SIFT description
85
13
27
7
...
121
128-dimensional luminance feature
64-dimensional chromaticity feature
BM25!
Topic 9068 "PUMA logo animal"
Topic 9048
1. video 65213
2. video 55238
...
1000. video 62469
Video ranked list for each topic
Frames
Video 76751

Figure 1: Instance search system for NTT-NII_1 run.

racy of the instance search task, we attempted to incorporate the BM25 ranking function(Robertson and Walker 1994), the state-of-the-art ranking function in the text retrieval field, in the video retrieval task. We extracted the keypoints from 21 instance topic images and video frames using the Harris-Laplace detector(Harris and Stephens 1988), and featured the keypoints on the basis of the 192-dimensional compact color SIFT method(Mikolajczyk and Schmid 2004)(Zhu and Sato 2012). The frame images were extracted from 76751 videos by the rate of one frame per second, and we obtained about a total of $740,000$ images. The first 128 dimensions in the compact color SIFT description correspond to the normal luminance SIFT and the latter 64 dimensions involve the chrominance information. The BM25 ranking function especially focused on the importance degrees of the keypoints and, using degrees, it ranked the videos according to the keypoint frequencies normalized by the video duration lengths.

1.2 Matching Keypoints

The extracted keypoints from instance topic images and video frames were featured by the 192-dimensional vectors and they were matched according to the cosine value between the two vectors. Then the keypoint pair showing the highest cosine value larger than 0.95 was considered as matched. The matching procedure was performed for every keypoint pair and it thus required huge computer power and most of the total computational time in our instance search system. The matched results are used afterward to calculate the factors in BM25 ranking function.

1.3 Inverse Frame Frequency(IFF)

The IFF estimates the importance degree of keypoints extracted from instance topic images. About $180,000$ different keypoints were extracted from 21 instance topic images and the IFFs are calculated using the video collection of instance search task as follows:

$$\mathrm{IFF}_i = \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \tag{1}$$

Here, IFF_i is the IFF weight of the ith keypoint. N is the total number of frame images in the video collection and n_i is the number of frame images containing the ith keypoint. Recall that n_i was already calculated by the matching keypoints procedure described above. The number 0.5 is added to avoid the zero division. IFF_i becomes high when n_i is small and so it indicates the occurrence tendency of each keypoint in the video collection. The frequently-appearing keypoints have low importance degrees because they are not useful in distinguishing and retrieving the videos of interest from the entire video collection.

1.4 BM25 Ranking Function with IFF Weights

The relevance score between the instance topic q and the video v was calculated by the following BM25I ranking function:

$$\mathrm{BM25I}(q,v) = \sum_{q_j} \mathrm{IFF}_j \frac{(k_1+1)\mathrm{KF}_j}{k_1\left(1-b_1+b_1\left(\mathrm{vl}/\mathrm{avvl}\right)\right)+\mathrm{KF}_j} \tag{2}$$

Here, q_j is the jth keypoint of q and the summation was performed over q_j. k_1 and b_1 are BM25 parameters and they are usually set as 1.2 and 0.75, which were determined from the past series of massive experimental results in the text retrieval field. Although the text retrieval and the video retrieval tasks are totally different, we chose 1.2 and 0.75 as well. KF_j is the frequency of q_j in the video v and the number is the sum of the frequency of q_j in each video frame. vl and avvl are the video length and the average video length, respectively. vl is the total number of keypoints extracted from the video frames, and it generally becomes large when the video duration is long. avvl is the average of vl in the entire video collection and the value was 47.11 this year. The videos were ranked in the decreasing order of BM25I from 1th to 1000th and the ranked list for each instance topic was evaluated using the search accuracy measure called mean average precision(MAP).

1.5 Mean Average Precision(MAP)

The MAP shown below is to measure the accuracy of the ranked list returned by the search system.

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{|R_q|}\sum_{k=1}^{1000} rel(q,k)\frac{c(q,k)}{k} \tag{3}$$

Here, $|Q|$ is the number of instance topics and it was 21 this year. $|R_q|$ is the number of correct videos regarding the instance topic q and the number differs for each instance topic. This year's average number of correct videos over 21 instance topics was 59. $rel(q,k)$ is an indicator function of relevance and irrelevance, and if the kth ranked video is relevant to q, $rel(q,k)$ is 1. On the other hand, if the kth ranked video is irrelevant to q, $rel(q,k)$ is 0. $c(q,k)$ is the accumulated number of correct videos retrieved upper kth rank. MAP is the average of average precision(AP) of the ranked list for each topic instance; that is, $\mathrm{AP}_q = \frac{1}{R_q}\sum_{k=1}^{1000} rel(q,k)\frac{c(q,k)}{k}$. The highest AP_q is 1.0 and the lowest value is 0.0.

2. INSTANCE SEARCH: NTT-NII_2 AND NTT-NII_3 RUNS

2.1 System Overview

Figure 2 shows an overview of NTT-NII_2 and NTT-NII_3 instance search systems. The differences from the NTT-NII_1 search system in Fig. 1 are the method of estimating keypoint importance degrees called RSJ(Robertson-Sparck-Jones) and the BM25 with the RSJ weight ranking function.

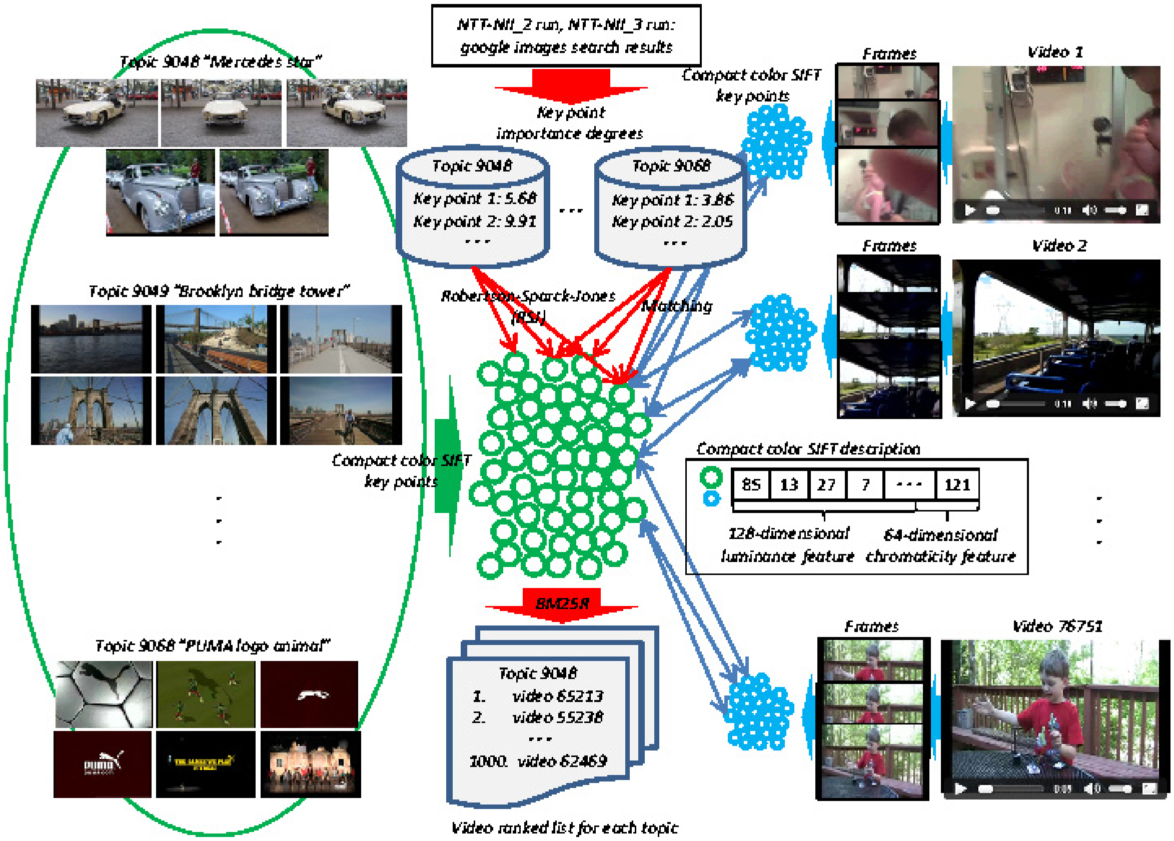


Figure 2: Instance search systems for NTT-NII_2 and NTT-NII_3 runs.

The IFF weights at BM25I ranking function in Eq. (2) were replaced by the RSJ weights. In calculating RSJ weights, the relevant images to each instance topic were required. We therefore collected them from the google images search results by inputting the keywords of the instance topics. The other parts of the search system were the same as those of the NTT-NII_1 run system.

2.2 Collecting Relevant Images to Instance Topics

For each instance topic, the topic keywords in Table 1 were available. We input the keywords to google images(http://images.google.com/) and regarded the top-ranked images as relevant ones to the instance topic. We collected about the top 50 ranked images for each instance topic, from which we extracted keypoints and featured them in the same manner as explained in Section 1.1. Each keypoint was compared with the keypoints extracted from the instance topic images and the keypoint pair showing the highest cosine value larger than α was considered as matched. Here, α is the threshold parameter and the settings of $\alpha = 0.95$ and $\alpha = 0.9$ correspond to NTT-NII_2 run

Table 1: Topic Keywords

Topic 9048	“Mercedes star”
Topic 9049	“Brooklyn bridge tower”
Topic 9050	“Eiffel tower”
⋮	⋮
Topic 9068	“PUMA logo animal”

and NTT-NII_3 run, respectively. $\alpha = 0.9$ yields softer keypoint matchings than those of $\alpha = 0.95$. The matching results were used to estimate the RSJ importance degree weights of keypoints.

2.3 Robertson-Sparck-Jones(RSJ)

RSJ weights are defined as follows:

$$\mathrm{RSJ}_{q,i} = \log\left(\frac{(r_{q,i}+0.5)(N-R_q-n_{q,i}+r_{q,i}+0.5)}{(n_{q,i}-r_{q,i}+0.5)(R-r_{q,i}+0.5)}\right) \tag{4}$$

Here, $r_{q,i}$ is the number of relevant images to the instance topic q containing the ith keypoint. N and R_q are the total number of images, $N = 50 \times 21$ in this case, and the total number of the relevant images to q. $n_{q,i}$ is the number of images containing the ith keypoint. $\mathrm{RSJ}_{q,i}$ becomes the simple IFF_i in Eq. (1) by setting $R = 0$ and $r_{q,i} = 0$; that is, no relevant images are available.

2.4 BM25 Ranking Function with RSJ Weights

The BM25 ranking function weighted by RSJ is formulated as follows:

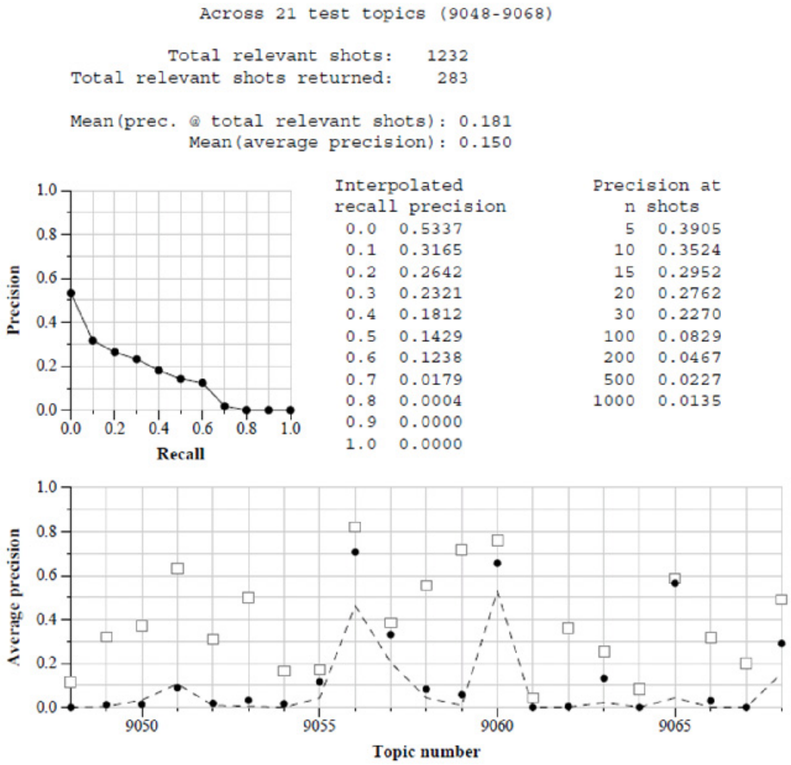
$$\mathrm{BM25R}(q,v) = \sum_{q_j} \mathrm{RSJ}_j \frac{(k_1+1)\mathrm{KF}_j}{k_1\left(1-b_1+b_1\left(\mathrm{vl}/\mathrm{avvl}\right)\right)+\mathrm{KF}_j} \tag{5}$$

Here, RSJ_j is the RSJ importance degree weight of the jth keypoint of q and the other notations and the parameter values are the same as those in Eq. (2). The videos were ranked in the decreasing order of BM25R and the ranked list for each instance topic was evaluated by using MAP.

3. EVALUATION RESULTS

3.1 NTT-NII_1 Run

Figure 3 shows the evaluation results for NTT-NII_1 run. The MAP was 0.15, the highest among our four submission runs. This run scored high APs for topics 9056 “Pantheon interior”, 9060 “Stephen Colbert”, and 9065 “Hagia Sophia interior”. The method for this run, BM25I, was capable of retrieving videos similar to the instance topic images and this feature led to the high APs. That is, for these topics, there were many relevant videos whose frame images were almost the



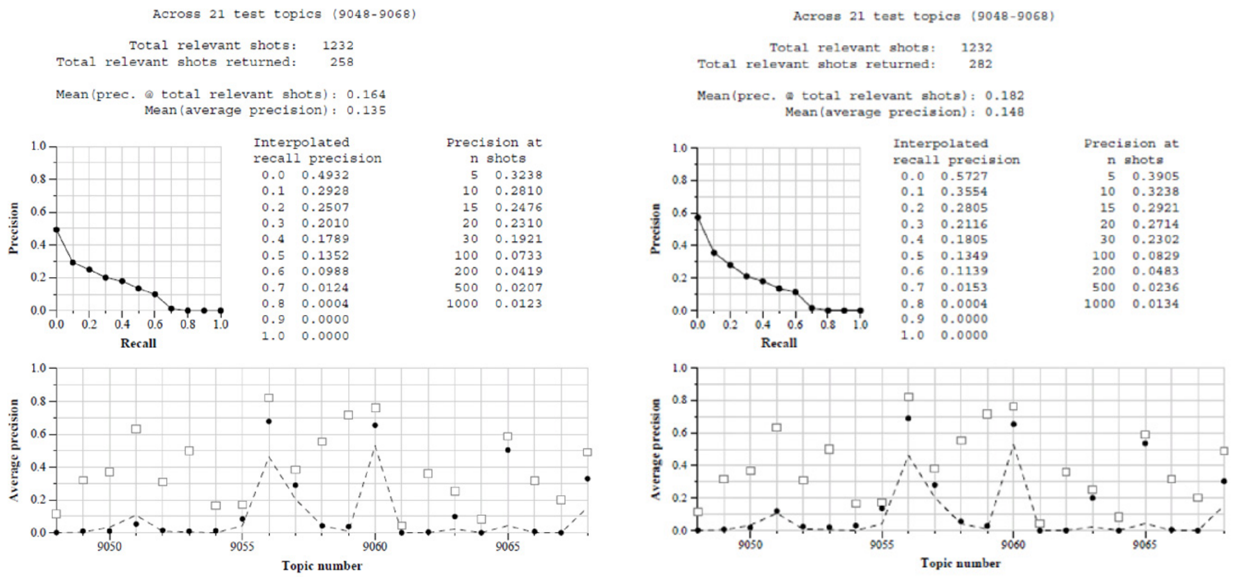
```
                 Across 21 test topics (9048-9068)

              Total relevant shots:   1232
        Total relevant shots returned:    283

        Mean(prec. @ total relevant shots): 0.181
                Mean(average precision): 0.150

                                Interpolated           Precision at
                                recall precision         n shots
                                 0.0  0.5337              5   0.3905
                                 0.1  0.3165             10   0.3524
                                 0.2  0.2642             15   0.2952
                                 0.3  0.2321             20   0.2762
                                 0.4  0.1812             30   0.2270
                                 0.5  0.1429            100   0.0829
                                 0.6  0.1238            200   0.0467
                                 0.7  0.0179            500   0.0227
                                 0.8  0.0004           1000   0.0135
                                 0.9  0.0000
                                 1.0  0.0000
```

Figure 3: NTT-NII_1 run results. Run score (dot) versus median (- -) versus best (box) by topic.

same as the topic images. As for topics 9055 "Sears/Willis Tower", 9057 "Leshan Giant Buddha", 9065 "Hagia Sophia interior", the method scored almost the best APs over all submission runs. This tendency seems also due to the method's capability of retrieving video frames similar to the topic images. On the other hand, the method failed to find the correct videos for topics such as 9048 "Mercedes star" and 9061 "Pepsi logo - circle". These topics were extremely difficult because the instance regions were small and the background information was not useful in the search. However, because we think tackling these topics is the mission of instance search, we will explore an effective solution in future work.

3.2 NTT-NII_2 and NTT-NII_3 Runs

Figure 4 shows the evaluation results for NTT-NII_2 and NTT-NII_3 runs. The MAPs of NTT-NII_2 and NTT-NII_3 runs were 0.135 and 0.148, respectively. The results indicate that the latter method, BM25R with $\alpha = 0.9$, is better. Because the 192-dimensional feature vectors were somewhat too detailed for the keypoints to be easily matched, the softer matching threshold was effective. Although the MAP and APs of the instance topics scored by the NTT-NII_3 run were almost



```
                Across 21 test topics (9048-9068)

            Total relevant shots:   1232
   Total relevant shots returned:    258

   Mean(prec. @ total relevant shots): 0.164
             Mean(average precision): 0.135

                        Interpolated              Precision at
                        recall precision            n shots
                         0.0  0.4932                 5  0.3238
                         0.1  0.2928                10  0.2810
                         0.2  0.2507                15  0.2476
                         0.3  0.2010                20  0.2310
                         0.4  0.1789                30  0.1921
                         0.5  0.1352               100  0.0733
                         0.6  0.0988               200  0.0419
                         0.7  0.0124               500  0.0207
                         0.8  0.0004              1000  0.0123
                         0.9  0.0000
                         1.0  0.0000
```

```
                Across 21 test topics (9048-9068)

            Total relevant shots:   1232
   Total relevant shots returned:    282

   Mean(prec. @ total relevant shots): 0.182
             Mean(average precision): 0.148

                        Interpolated              Precision at
                        recall precision            n shots
                         0.0  0.5727                 5  0.3905
                         0.1  0.3554                10  0.3238
                         0.2  0.2805                15  0.2921
                         0.3  0.2116                20  0.2714
                         0.4  0.1805                30  0.2302
                         0.5  0.1349               100  0.0829
                         0.6  0.1139               200  0.0483
                         0.7  0.0153               500  0.0236
                         0.8  0.0004              1000  0.0134
                         0.9  0.0000
                         1.0  0.0000
```

Figure 4: (Left) NTT-NII_2 run results. (Right) NTT-NII_3 run results. Run score (dot) versus median (- -) versus best (box) by topic.

identical to those for the NTT-NII_1 run, BM25R with $\alpha < 0.9$ might further improve the search accuracy. Moreover, the larger number of relevant images might lead to better estimation of keypoint importance degrees. An investigation of these parameter effects remains as future work.

4. MULTIMEDIA EVENT DETECTION TASK

4.1 MED System Overview

Our MED system is based on SVM classifiers defined on a semantic space that is obtained by the canonical correlation analysis (CCA) of image and text data. Fig. 5 shows a schematic view of our MED system.

4.2 Image and Text Features

For all processes of training, metadata generation, and event agent generation, images are captured from given video clips every ten seconds and a vector is generated for each image by bag-of-features approaches. We implemented two kinds of features: SIFT descriptors (Lowe 2004) and RGB colors. They are used for separate runs. Finally, a 1,000 dimensional image vector is obtained for each image by calculating a normalized histogram, where frequency bins are represented by centroids calculated by the k-means clustering algorithm. To calculate k-means centroids, we used one million randomly sampled image vectors, which were extracted from video clips in the event kit.

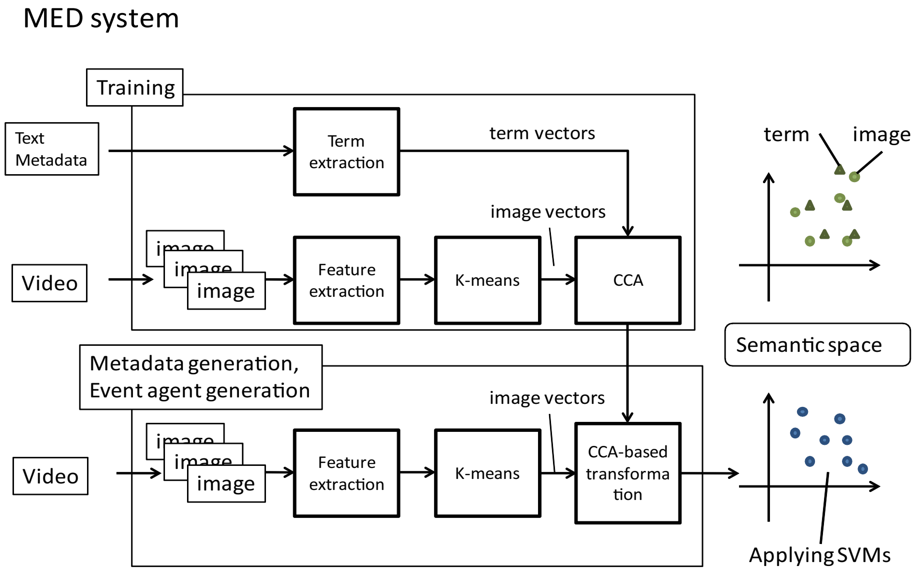


Figure 5: A schematic view of our MED system. In the training phase, a semantic space is generated using text and video data. In the metadata generation and event agent generation phase, image vectors are projected onto the semantic space.

In the training phase, a semantic space is determined using video clips and text-based Event Metadata provided as "*_JudgementMD.csv" files. We extracted terms appearing in four entries of the metadata, which are "Synopsis," "Scene," "Object," and "Activity". After eliminating "stop words", which are 488 very frequent terms such as pronouns and prepositions and applying Porter's stemming algorithm, a binary term vector, which represents existence of each term, is generated corresponding to each video clip. We obtained a 3,191 dimensional vector for each of 3,336 video clips from the provided event kit data (EVENTS_20120227_JudgementMD.csv) . In the proposed system, we do not use term frequencies because the same words tend to be used redundantly in multiple entries of the Event Metadata.

4.3 Semantic Space Generation Using CCA

The canonical correlation analysis (CCA) is a statistical method used to analyse two different sets of vectors. It has also been applied to image retrieval tasks with multiple data types such as image and text data (Hardoon, Szedmak and Shawe-Tayer 2004). The CCA determines a couple of linear transformations that yields bases where samples from two data sets are well-correlated. Applying the CCA to image and term vectors, it is expected that image vector elements that are not related to meanings, which are represented by text vectors, can be eliminated. In the implementation, the identical term vectors are assigned to image vectors extracted from the same video clip. We call the obtained space “semantic space” in this system (Fig. 5).

To eliminate elements with low correlation and to avoid over-fittings to training data, we chose bases with correlation coefficients from 0.55 to 0.95 on the semantic space. Eventually, each captured image is represented as a vector on a semantic space calculated by the above-described procedures, in both the metadata generation and event agent generation phase.

4.4 Detection Based on SVM Classifiers

In the event agent generation phase, we trained a classifier for each event using the support vector machine (SVM). We used the LIBSVM package (Chang and Lin 2011). As detection scores, we adopt probability estimation outputs of the LIBSVM, which approximate probability distributions by the sigmoid function of SVM's discriminant functions. For SVM training, we used a second-order polynomial kernel and set the parameter c at 0.1.

Finally, a detection score is determined by the maximum probability value among the captured images for each test video clip[1]. We divided video clips into two parts. One is for SVM training and the other is for validation where detection thresholds are determined. We simply used break-even-points of recall and precision values as detection thresholds.

4.5 Submitted Runs

We executed four runs for the Pre-Specified event task and compared two image features (the SIFT descriptors and color histograms) and effects of the CCA with text vectors. The runs consist of combinations of image features and presence or absence of the CCA. In the absence case, image vectors are directly used, instead of a semantic space. Compositions of runs are as follows:

p-baseline_1 (PS) SIFT descriptors are used for features and the CCA with text data is *not* used.

c-contrast_1 (PS) SIFT descriptors are used for features and the CCA with text data is used.

c-contrast_2 (PS) Color histograms are used for features and the CCA with text data is *not* used.

c-contrast_3 (PS) Color histograms are used for features and the CCA with text data is used.

For the Ad Hoc event task, we submitted two runs only using SIFT descriptors as image features. p-baseline_1 (AH) and c-contrast_1 (AH) denote runs with and without the CCA, respectively.

Table 2 shows the evaluation results for submitted runs. It is not easy to compare the detection performance directly because the proportions of P_{Fa} to P_{Miss} is largely different among runs.

[1] For the Pre-Specified event task, we used video clips in the event kit for negative samples because we did not notice the FAQ #3 description on the MED'12 web page. For the Ad Hoc event task, we appropriately prepared negative samples from MED'11 clips.

Table 2: Evaluation results for submitted runs. NDC, P_{Fa}, and P_{Miss} denote the normalized detection cost, false alarm ratio, and miss detection ratio, respectively.

Run	Image Feature	CCA	NDC	P_{Fa}	P_{Miss}
p-baseline_1 (PS)	SIFT	*not* used	2.3070	0.1149	0.8721
c-contrast_1 (PS)	Color hist.	*not* used	1.1456	0.0153	0.9543
c-contrast_2 (PS)	SIFT	used	9.3010	0.7351	0.1210
c-contrast_3 (PS)	Color hist.	used	2.3930	0.1281	0.7933
p-baseline_1 (AH)	SIFT	*not* used	8.7419	0.6702	0.3724
c-contrast_1 (AH)	SIFT	used	1.0000	0.0000	1.0000

Inappropriate setting of detection thresholds would cause this property. This is one reason for the very high normalized detection cost (NDC) values.

Comparing the NDC values, we see that semantic space generation using the CCA with text data did not work effectively. One conceivable reason is too little text information. Actually, the average number of non-zero elements of term vectors is about 8.2 and the mode is 4. Other resources that combine text and image (or video) data, in addition to the provided Event Metadata, should be used to construct appropriate semantic spaces.

5. CONCLUDING REMARKS

Our findings from this year's instance search experiments are summarized as follows:

1. BM25 is promising as the baseline video ranking approach for the instance search task.

2. Softer keypoint matching is appropriate for the estimation of keypoint importance degrees.

Future work includes an experimental investigation of the BM25R ranking method with a much softer matching threshold. We will also enhance the number of relevant images for each topic and check the impact. As for the BM25I ranking function, we will take into account the mask images and give more weights to the keypoints inside the instance regions specified by the masks.

As for this year's multimedia event detection, we constructed a MED system based on semantic space generation from image and text features using the CCA and detection using SVM classifiers. However, the evaluation results for MED'12 show that image and text data combination does not work effectively with the provided Event Metadata.

REFERENCES

Chang, C. C., and Lin, C.-J. (2011), "LIBSVM : a library for support vector machines," *ACM Trans.actions on Intelligent Systems and Technology*, 2(3), 27:1–27:27.

Hardoon, D. R., Szedmak, S., and Shawe-Tayer, J. (2004), "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computaion*, 16, 2639–2664.

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

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

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

Robertson, S. E., and Walker, S. (1994), "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval.," *Proc. of SIGIR'04*, pp. 232–241.

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