TRECVID 2010 Known-item Search by NUS

Xiangyu Chen, Jin Yuan, Liqiang Nie, Zheng-Jun Zha, Shuicheng Yan
Tat-Seng Chua

National University of Singapore, Singapore

Abstract. This paper describes our system for auto search and interactive search in the known-item search (KIS) task in TRECVID 2010. KIS task aims to find an unique video answer for each text query. The shift from traditional video search has prompted a series of challenges in processing and searching techniques that developed over the past few years. For the automatic search task, our **VisionGo** system performs query expansion and analysis, then employs multi-modality features including metadata, automatic speech recognition (ASR) and high level feature (HLF) to retrieve a ranked list of results deemed most relevant to the text-only query. To further improve the search performance, we crawl an extension set of tags from Youtube to supplement to TRECVID metadata. For interactive search task, we propose a new feedback scheme based on both related samples and exclusive negative samples to boost the search performance. To accomplish this, we introduce three enhancements to our **VisioGo** system: a) related sample feedback algorithm that allows users to indicate related (but not relevant) shots to the query; b) exclusive negative sample selection approach; and c) clustered shot-icons for efficiently representing the whole content of the video. Results from TRECVID 2010 video test set indicate that the enhancements are effective.

1 Introduction

The known-item search (KIS) task in TRECVID poses a new challenge for video retrieval as it requires the system to locate an unique video answer for the guided text-only query, which models real-world scenario. According to this extreme task of KIS, we participate in both the auto search and interactive search, and focus on providing effective video retrieval based on our video search platform-**VisionGo**. Three essential features are embedded in our **VisionGo** system [1]: a) a well-performed automatic search engine; b) the proposed related samples and exclusive negative samples based feedback technique; and c) an efficient user interface (UI) for good interaction and efficient visualization.

For auto search, our **VisionGo** system performs query analysis and multi-modality fusion based on our previous work [1, 2]. In order to supply a high-quality initial result list for the next interactive search, our auto search engine focuses on the incorporation of various features like ASR, HLF and the metadata provided by TRECVID. To further improve the search performance, an extension set of tags is crawled from Youtube based on the title of each video of the

TRECVID dataset. These Youtube tags are indexed to supplement our auto search engine.

For interactive search, the main challenge is that there is insufficient relevant samples to help in finding the desired answer for the text-only query. In order to overcome this problem, we propose a novel framework in **VisionGo** interactive search engine. It utilizes the video results from the automatic search as initial retrieval set for users' feedback. The users then screen the returned video results and select a set of related shots from different videos to indicate whether they are indeed related or not through an efficient user interface. The UI of **VisionGo** provides an intuitive visualization of the visual content of the videos to facilitate users' annotation efforts. The interface allows the users to see a dynamic series of clustered shot-icons instead of a single key frame to represent the full content of each video. In the feedback process, we design a novel feedback algorithm based on both the related and exclusive negative samples. The purpose is to provide enough related samples [3] to the complex text query, and automatically provide exclusive negative samples to refine the query learning and improve the subsequent searches.

2 Automatic Search Task

2.1 Overview of the Auto Search Process

In the past research of video retrieval,ASR [4] and HLF [5] are found to important to enhance the performance. However, in the KIS of this year, the metadata is the most effective textual modality while ASR is likely to play a complementary role. Our focus this year is therefore on effective video retrieval employing the multi-modality fusion with Metadata, ASR and HLF [6–15]. To further improve the auto search performance, we crawl an expansion set of tags from the Youtube website. The tag set consists of 8,383 subsets of Youtube tags. Each subset is downloaded according to the title of each video in the test dataset of TRECVID. For video transcript, we adopt the ASR result donated by LIMSI and Vecsys Research [16].

For the indexing of our **VisonGo** system, we adopt the Lucene index technology [17] to obtain two main indexes: *Meta Index* and *Youtube Index*. ASR information is induced in these two indexes as complementary part. The auto retrieval starts with the user's text-only query, which describes the events happened in the video. Our system then performs query preprocessing and uses Lucene retrieval technology, which adopts tf-idf weighting scheme, to obtain an initial ranked list. Finally, we employ two kinds of modality fusions (*Youtube+Metadata* and *HLF+Metadata*) to derive the final list of results deemed most relevant to the query. The detail process of our auto search task is explained in Figure 1.

2.2 Query Analysis

The text-only queries of the KIS task are very diverse and complex. For example, "Find the video of a Asian family visiting a village of thatch roof huts showing

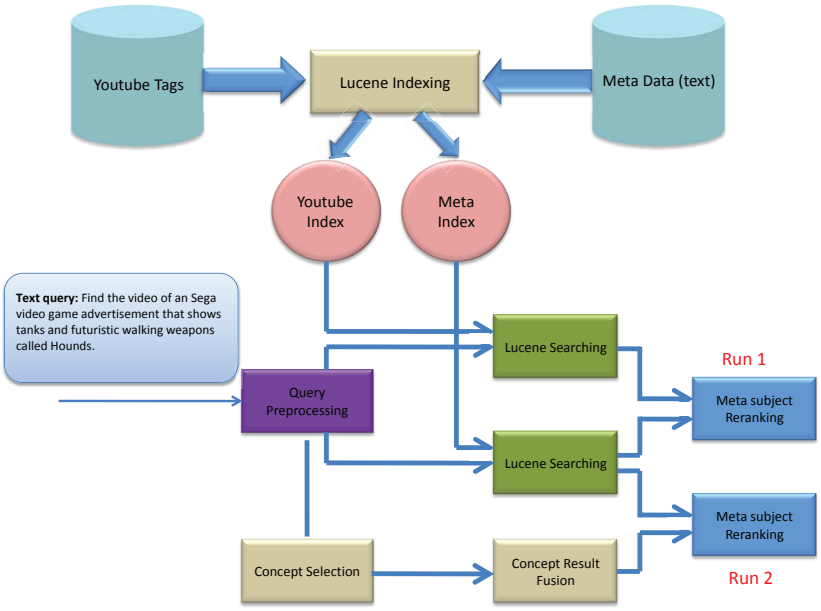

Fig. 1. Overview of Automatic Search Task

two girls with white shirts and a woman in red shorts entering several huts with a man with black hair doing the commentary."This kind of query describes the main events that happened in the whole video, in which some abstract words are beyond the range of 130 concepts in the semantic indexing task of TRECVID. So query analysis and expansion are important in KIS task as they help to understand the users' intention and retrieve the most relevant/related answers.

As observed, the metadata play a key role in this year video retrieval. However, the mapping between the query and the metadata is a challenging problem because some key words in query might not be found in the metadata OR inversely some abstract key words in the metadata are not in the list of 130 concepts. This motivates us to gather extra context for each query. We therefore adopt query expansion by generating additional relevant key words for each query in the following two steps (similar as [18]):

(a) Use the complex query to retrieve relevant videos from Youtube and collect the tags/comments of these relevant videos;
(b) Extract terms from the set of relevant tags/comments, which have high mutual information (MI) with the key words of the query.

In addition, HLF is also useful to the query in terms of visual requirements. It is clear that the kind of visual-oriented query such as "Find the video of bald, shirtless man showing pictures of his home full of clutter and wearing headphone" requires some visual cues in the shot which cannot be achieved with

only text features. Our System approach this by employing morphological analysis followed by selective expansion using the WordNet [19] on both the feature descriptions of HLFs and KIS's queries [5]. The stronger the match between the HLF descriptions and the query, the more important the HLF is to the query.

3 Interactive Search Task

For interactive search task in KIS, we emphasize not only on retrieval performance but also on enhancing user's annotation efforts through advanced visualizations for a efficient user interface. Therefore our **VisionGo** system focuses on maximizing the human annotator effort through the use of: (1) efficient User Interface (UI); (2) the proposed feedback method based on both related samples and exclusive negative samples; and (3) clustered shot icons for fast previewing of the main content of the videos. The system first provides the user with the results from the automatic search. The user can make use of UI to refine the search results with the proposed new feedback algorithm. During the process of interactive search, users are able to see the dynamic series of shot-icons instead of a single keyframe.

3.1 User Interface

The KIS task has posed a challenging problem, which require the system to locate an unique true answer according to the complex query. This kind of query not only describes the whole events contained in the video but also includes some audio cues as well as visual cues. For example,"Find the video about the cost of drug, featuring a man in glasses at a kitchen table, a video of Bush, and a sign saying Canada." This query describes the main events happened in this video, which cover the visual contents of the video. There is still audio search cue such as "a sign saying Canada". Hence the key in interactive search is how to facilitate the user to efficiently go through the content of the whole video and maximize user's annotation efforts. The UI of VisionGo is designed for fast mouse-clicking with quick preview of the main content of each video in the ranked list. A sample of our efficient UI is shown in Figure 2.

The UI will display the ranked results of auto search through 5×8 grid in the "Results Show Area". This maximizes the speed of annotation of user. After going through the dynamic series of shot-icons for each video in this area, the user can click the set of related videos. Based on the selecting, a series shot-icons of the clicked videos (here we constrain each video to have at most 10-15 shot-icons) will be show at the bottom of UI in the "Related Samples Selection Area". The user can then select the related shot-icons from different videos to compose a new visual video query in the "Feedback Area", including the ASR, HLF and metadata from each of the selected shots. Through the new video query, the interactive search engine performs the proposed feedback based on the selected related samples and automatically learned the exclusive negative

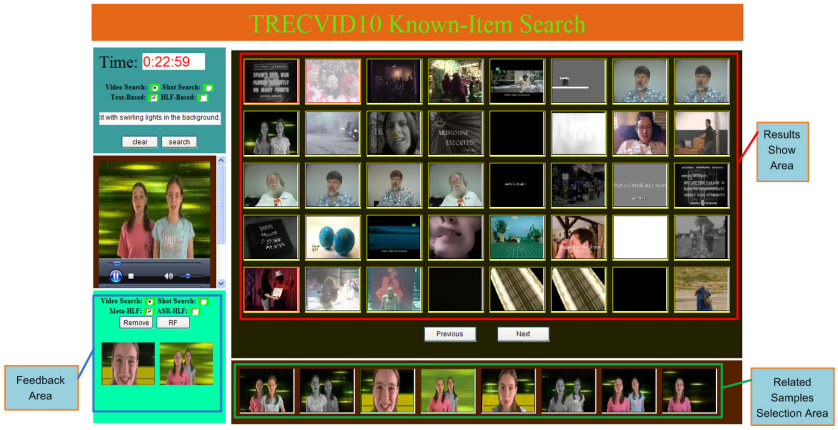

Fig. 2. Efficient User Interface

samples. Finally, the new results will be shown in the "Results Show Area". The users can repeat the above search process until the desired result has been found.

In addition, as observed in the above query example, there are several audio search cues in many queries such as "a sign saying Canada". For the word "Canada", the Luence search engine may easily find it in the metadata or ASR of the videos; but for the words "a sign saying", the metadata or ASR may be useless. This motivate us to embed a play function in the UI to play the videos from a clustered visual shot-icons when the user wants to find some special cues not present in the text modalities.

3.2 Proposed New Feedback Scheme

Related Samples Strategy One of the main challenges in the TRECVID known-item search is the insufficient relevant sample according to the task queries, which consists of complex semantics. To address this problem, we utilize the related shot samples to do feedback and learn the desired complex query [3]. Here the related shot samples refer to those shot segments from the selected videos, that are irrelevant to the whole text query but relevant to some of the related concepts of the query. From our observation in this year's TRECVID videos, the relevant videos are really rare, but the related shots are usually availabe and easy for users to annotate. In addition, we learn a detector for the query by simultaneously leveraging the related concept detectors, as well as user's feedbacks including related positive samples and exclusive negative samples. The exclusive concepts set can be automatically learned from the training data of TRECVID 2010.

Given a complex query Q, a set of related concept detectors $\{f_k\}_{k=1}^{K}$ can be learned from the text-based detector selection scheme [20]. Let $D^t = \{s_i, y_i\}_{i=1}^{N_t}$ denote the labeled samples in feedback iteration t, where s_i is the text or visual feature vector of shot sample i, and $y_i \in [0, 1]$ is the label derived from user's

feedback. Different from [3], y_i indicates whether s_i is related ($y_i \in 0, 1$) or unrelated ($y_i = 0$) to the overall query. For a related shot sample, y_i is estimated to measure the "related strength" of s_i with respect to query Q [3]. For the unrelated samples, different from the traditional algorithm (consider all the unlabeled samples as negative), we learn the exclusive negative subset as negative samples, which will be detailed in next subsection.

For each iteration, we can learn a query detector $f^t(s)$ from D^t:

$$f^t(s) = \eta^t \sum_{k=1}^{K} d_k^t f_k(s) + \frac{1}{t-1} \sum_{l=1}^{t-1} \beta_l^t \Delta f^l(x) + \Delta f^t(s), \tag{1}$$

where $\{d_k^t\}_{k=1}^K$ are the weights of concept detectors at iteration t; $\{\beta_l^t\}_{l=1}^{t-1}$ are the weights of the previous delta detectors; and η^t is a trade-off parameter which balances the concept detectors and the delta detectors. The solution can be obtained by solving the optimization problem in [3].

In this known-item search task, since there is only one target answer, we adopt a fusion strategy at each iteration to learn video detector by combining the shot detector scores. The fusion algorithm is as below:

$$F^t(v_j) = \frac{1}{N_{v_j}} \sum_{p=1}^{N_{v_j}} f^t(s_p), \tag{2}$$

where s_p is the shot of video v_j and N_{v_j} is the number of shots of v_j.

As mentioned before, the user might select and combine the related shot-icons from different videos to form a new visual video query in the "Feedback Area", and fusing the ASR, HLF and metadata from each of the selected shots. The fusion process can be shown in Figure 3

From the observation from the content of meta data files of TRECVID, for each video, we only consider the "title", "subject"and "description"parts in the metadata file. For retrieval, we first utilize Lucence to obtain the match score between the text query and each of the three parts; and then fuse these three scores with the optimal weights. During feedback, for each related shot, we get the correspond shot metadata through following steps:

(a) Combine the three parts ("title", "subject"and "description") in metadata file of the video as a fusion text file and utilize it based on the sentence unit;

(b) Learn the matching score between the text query and each sentence of the combined metadata file of each shot, the top 2 sentences will be considered as the metadata of each related shot.

Finally, through above strategies, we obtain the ASR, metadata and HLF detectors of each of the selected related shots. After converting them from the shot level to video level, we obtain a visual video query that includes metadata, ASR and HLF.

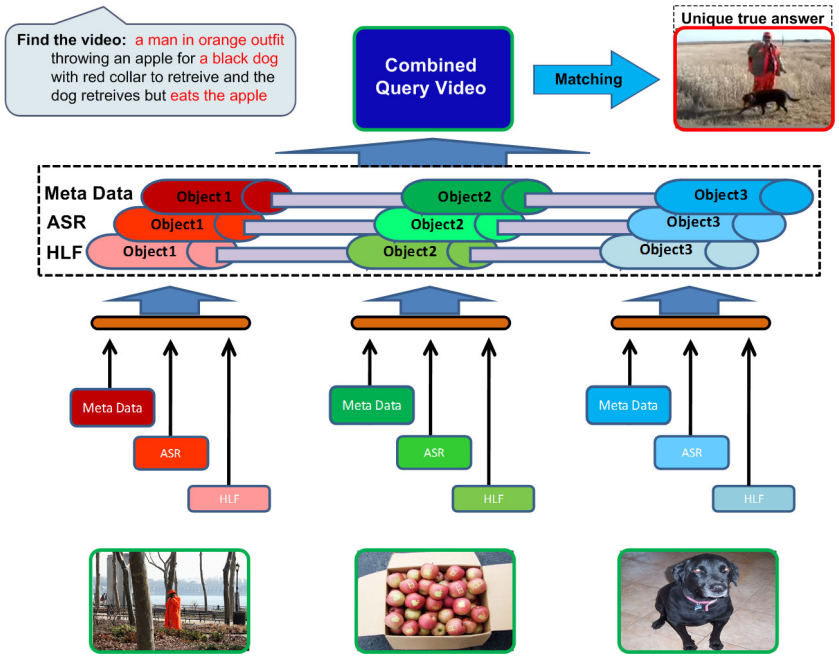

Fig. 3. The Fusion Process of Related Samples

Exclusive Negative Samples Selection Most of the interactive retrieval engines in TRECVID require the users to only indicate if the shots are positive for the given query [21, 22]and the rest of the un-tag shots are automatically taken as negative, which make the training process unbalance. This year, we adopt a novel strategy for selecting the exclusive negative sample from the un-tagged shots/videos corresponding to the indicated related samples. These samples are automatically selected based on the learned exclusive concept sets and are used as negative training sample for the feedback. Below is the brief introduction of this strategy. The examples of the learned exclusive concept subsets are shown in Figure 4. It should be noted that we learn the "exclusive" subsets only from the distribution statistic of the training data.

For learning the exclusive concept subsets, we adopt the graph shift method [23] based on weighted concept graph. Given a weighted concept graph $G =< V, E >$, the node set $V := \{1, 2, ..., n\}$ (n is the number of all the concepts in the dataset) and the edge set $E = V \times V$, we learn the exclusive relationships of the concepts based on the graph topology. Here the concept graph can be represented by a weight matrix D. In this matrix, the element d_{ij} will be assigned to a large value if concept i and concept j do not simultaneously appear in any training key frame or just in rare key frames (we will give a threshold), which depend the distribution of the training dataset; otherwise d_{ij} is equal to zero (the setting is similar as [24]). We can obtain the dense subgraph of D.

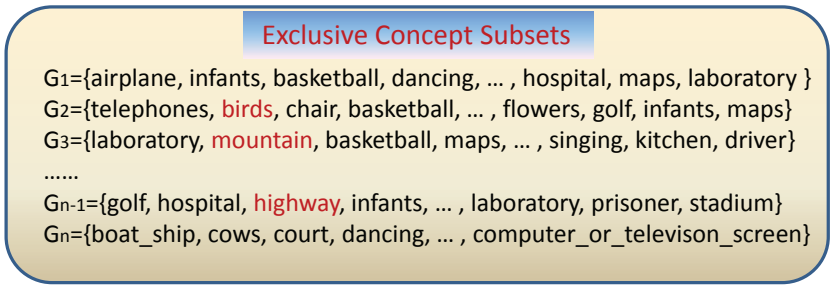

Fig. 4. Exclusive Concept Subsets learned from Training Data

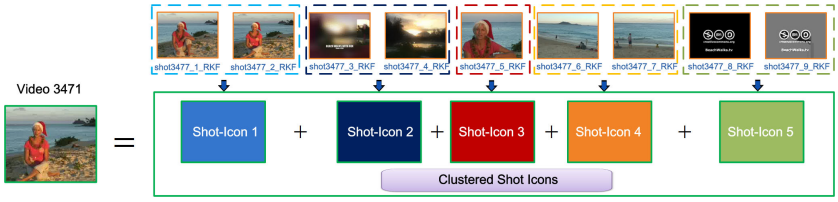

Fig. 5. The Clustering Process for Visualizing Video

The next stage is how to obtain the exclusive negative samples. In our system, if the selected related samples contain the concepts: "birds", "mountain", "highway", then the exclusive negative set for the query is $G_e = (G_2 \cup G_3 \cup G_{n-1}) \setminus \{\text{“}birds\text{”}, \text{“}mountain\text{”}, \text{“}highway\text{”}\}$. Finally, the negative samples are the shots or videos contain at least one concept in G_e.

3.3 Clustered Shot-icons

From the analysis of all queries of KIS, we find that many queries of this year cover many diverse events took place in videos. So it is necessary to design some thing new like Motion-icons [1] than just a set of keyframes to represent the content of videos. Unlike the traditional way of displaying the static keyframe for each video or shot, we adopt the dynamic series of clustered shot-icons to represent the full content of each video. Each clustered shot-icon is obtained through clustering the key frames of all shots, where the key frames with similar visual features will be collected as a cluster set. In each of the cluster set, we select at most five key frames to dynamically represent the cluster. We then combine the obtained clustered shot-icons as a summarized clip to visualize the video content in chronological order. The whole process is shown in Figure 5.

3.4 Fusion of Two Kinds of HLF

In past years, high-level features (HLF) [25–30] have been well studied to incorporate textual features to improve the retrieval performance. This year we

utilize the fusion detector scores of Columbia University [31] and label propagation results from [32] to boost the interactive search performance.

A concept detector in [33, 34] is trained for each query using SVM and kNN, before selecting the concepts using visual features and text descriptions. For the multi-label propagation algorithm, we extract multiple types of visual features from each key frame: 225-D blockwise color moments, 128-D wavelet texture and 75-D edge direction histogram. The setting is similar to [32]. We then make use of the available collaborative annotation results for training and learn the 130 detectors for each of the key-frames of the testing data provided by TRECVID [35]. The advantages of the multi-label propagation are two-fold: (1) the computation cost is lower than that of SVM and kNN methods, where this method only takes about 32 hours to obtain the detection scores of the whole test dataset of KIS; and (2) the learned concept scores are robust to noises because this approach is based on the hashing-accelerated l_1-graph construction and KL-divergence oriented optimization as validated in [32].

After the linear combination of these two kinds of concept detections, we also give the fusion HLF detetion scores an empirical threshold, which makes the detections accepting above 0.25.

4 Results and Analysis

We submitted four runs in KIS for auto search task and interactive search task, which are as follows:

- **Run1:** Fully automatic search employing multi-modality features including metadata, ASR and Youtube tags;
- **Run2:** Fully automatic search employing multi-modality features including metadata, ASR and HLF;
- **Run3:** Interactive search employing multi-modality features including metadata, ASR, HLF and Youtube tags;
- **Run4:** Interactive search employing multi-modality features including metadata, ASR and HLF;

The performance for these submitted runs are shown in Table 1. From the results, we can find that the text futures are important to locate the true answer video in both auto search and interactive search.

Table 1. Performance of All Runs in Our VisionGo System

Run ID	Mean Inverted Rank	Mean Elapsed Time
Run 1	0.215	0.021
Run 2	0.217	0.021
Run 3	0.682	2.577
Run 4	0.682	2.779

For auto search, the HLF is not effective because the mapping between the query and the 130 pre-defined concepts does not work well. Some terms used in queries are too abstract and specific to be mapped to the correct concepts. This year we tried to utilize the downloaded Youtube data to give more correct mapping to boost the auto performance. However, the tags in Youtube are also diverse as the terms in metadata in TRECVID. So the performance of Run 1 was as bad that of Run2.

For interactive search, in Run 3 and Run 4, we took adavantage of the multi modal features available to boost the search performance. The difference between these two runs is that the former makes use of the downloaded Youtube tags during fusion. The mean inverted ranks of Run3 and Run 4 are the same (0.682), which is the top two performance in all interactive search participants. This validate our proposed feedback scheme based on both related samples and exclusive negative samples. The feedback refines and learns the visual video query which consists of the information of metadata, ASR, HLF and Youtube tags during the whole process of interactive search.

5 Conclusion

This paper introduced and discussed the details of our participation in TRECVID 2010 Known-item search. We described the framework and techniques we employed for automatic search and interactive search tasks.

For auto search task, our **VisionGo** system focuses on the the incorporation of various features like ASR, HLF and the metadata provided by TRECVID. To further improve the search performance, an extension tag data set are crawled from Youtube and are indexed in our auto search engine. For interactive search, since the main challenge is that there is insufficient relevant samples for the desired query. In order to overcome this problem, we proposed a novel framework in VisionGo to boost interactive search performance through the use of: (1) efficient UI; (2) the proposed feedback method based on both related samples and exclusive negative samples; and (3) clustered shot icons for fast previewing main content of the videos. The evaluation results indicate that Our **VisionGo** system are effective. Especially in the interactive search, our system has good overall performance.

6 Acknowledgments

This research was supported by NRF (National Research Foundation of Singapore) Research Grant 252-300-001-490 under the NExT Search Center and in part by the NRF/IDM Program under Research Grant NRF2008IDM-IDM004-029.

References

1. Zheng, Y.-T., Neo, S.-Y., Chen, X., Chua, T.-S.: Visiongo: towards true interactivity. In: Proceedings of ACM Conference on Image and Video Retrieal (CIVR). (2009)
2. Chua, T.-S., Neo, S.-Y., Li, K.-Y., Wang, G., Shi, R., Zhao, M., Xu, H.: Trecvid 2004 search and feature extraction task by nus pris. In: TREC Video Retrieval Evaluation Online Proceedings. (2004)
3. Yuan, J., Zha, Z.-J., Zhao, Z., Zhou, X., Chua, T.-S.: Utilizing related samples to learn complex queries in interactive concept-based video search. In: Proceedings of CIVR. (2010)
4. Huijbregts, M., Ordelman, R., de Jong, F.: Annotation of heterogeneous multimedia content using automatic speech recognition. In: Proceedings of International Conference on Semantic and digital Media Technologies. (2007)
5. Neo, S.-Y., Zhao, J., Kan, M.-Y., Chua, T.-S.: Video retrieval using high-level features: Exploiting query-matching and confidence-based weighting. In: Proceedings of CIVR. (2006)
6. Chua, T.-S., Neo, S.-Y., Zheng, Y.-T., Goh, H.-K., Zhang, X.: Trecvid 2007 search tasks by nus-ict. In: TREC Video Retrieval Evaluation Online Proceedings. (2007)
7. Zha, Z.-J., Yang, L., Mei, T., Wang, M., Wang, Z.: Visual query suggestion. In: Proceedings of ACM Multimedia. (2009)
8. Cao, J., Zhang, Y.D., Feng, B.L., Bao, L., Pang, L., Li, J.-T., Gao, K., Wu, X., Xie, H.-T., Zhang, W., Mao, Z.D.: Trecvid 2009 of mcg-ict-cas. In: TREC Video Retrieval Evaluation Online Proceedings. (2009)
9. Yanagawa, A., Chang, S.-F., Kennedy, L., Hsu, W.: Columbia university's baseline detectors for 374 lscom semantic visual concepts. Columbia University ADVENT Technical Report #222-2006-8 (2007)
10. Zha, Z.-J., Mei, T., Wang, J., Wang, Z., Hua, X.-S.: Graph-based semi-supervised learning with multiple labels. Journal of Visual Communication and Image Representation **20** (2009) 97–103
11. Zha, Z.-J., Mei, T., Wang, Z., Hua, X.-S.: Building a comprehensive ontology to refine video concept detection. In: Multimedia Information Retrieval. (2007)
12. Wang, M., Hua, X.-S., Tang, J., Hong, R.: Beyond distance measurement: Constructing neighborhood similarity for video annotation. IEEE Transactions on Multimedia **11** (2009) 465–476
13. Wang, M., Hua, X.-S., Hong, R., Tang, J., Qi, G.-J., Song, Y.: Unified video annotation via multi-graph learning. IEEE Transactions on Circuits and Systems for Video Technology **19** (2009) 733–746
14. Zheng, Y., Zhao, M., Neo, S.Y., Chua, T.S., Tian, Q.: Visual synset: Towards a higher-level visual representation. In: Computer Vision and Pattern Recognition. (2008)
15. Tang, J., Hua, X.-S., Qi, G.-J., Song, Y., Wu, X.: Video annotation based on kernel linear neighborhood propagation. IEEE Transactions on Multimedia **10** (2008) 620–628
16. Gauvain, J., Lamel, L., Adda, G.: The limisi broadcast news transcription system. Speech Communication (2002)
17. McCandless, M., Hatcher, E., Gospodnetic, O.: Lucene in Action. Manning Publication (2010)
18. Neo, S.-Y., Zheng, Y.-T., Goh, H.K., Chua, T.-S.: News video retrieval using implicit event semantics. In: International Conference on Multimedia & Expo. (2007)

19. Fellbaum, C.: WordNet: an electronic lexical database. The MIT press (1998)
20. Chang, S.-F., Hsu, W., Jiang, W., Kennedy, L.S., Xu, D., Yanagawa, A., Zavesky, E.: Columbia university trecvid-2006 video search and high-levelfeature extraction. In: TRECVID Workshop. (2006)
21. Luan, H.-B., Neo, S.-Y., Goh, H.K., Zhang, Y.-D., Lin, S.X., Chua, T.-S.: Segregated feedback with performance-based adaptive sampling for interactive news video retrieval. In: Proceedings of ACM Multimedia. (2007) 253–262
22. Luan, H.-B., Zheng, Y.-T., Neo, S.-Y., Zhang, Y., Lin, S., Chua, T.-S.: Adaptive multiple feedback strategies for interactive video search. In: Proceedings of CIVR. (2004)
23. Liu, H., Yan, S.: Robust graph mode seeking by graph shift. In: International Conference on Machine Learning. (2010)
24. Chen, Q., Song, Z., Liu, S., Chen, X., Yuan, X., Chua, T.-S., Yan, S., Hua, Y., Huang, Z., Shen, S.: Boosting classification with exclusive context. In: VOC Workshop: http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/workshop/nuspsl.pdf. (2010)
25. Mei, T., Zha, Z.-J., Liu, Y., Wang, M., Qi, G.-J., Tian, X., Wang, J., Yang, L., Hua, X.: Msra att trecvid 2008: High-level feature extraction and automatic search. In: TREC Video Retrieval Evaluation Online Proceedings. (2008)
26. Mei, T., Hua, X.-S., Lai, W., Yang, L., Zha, Z.-J., Liu, Y., Gu, Z., Qi, G.-J., Wang, M., Tang, J., Yuan, X., Lu, Z., Liu, J.: Msra-ustc-sjtu at trecvid 2007: High-level feature extraction and search. In: TREC Video Retrieval Evaluation Online Proceedings. (2007)
27. Tang, S., Li, J.-T., Li, M., Xie, C., Liu, Y.Z., Tao, K., Xu, S.X.: Trecvid 2008 high-level feature extraction by mcg-ict-cas. In: TRECVID Workshop. (2008)
28. Gao, Y., Dai, Q.: Clip-based video summarization and ranking. In: Proceedings of CIVR. (2008)
29. Gao, Y., Wang, W.B., Yong, J.H., Gu, H.J.: Dynamic video summarization using two-level redundancy detection. Multimedia Tools and Applications **42** (2009) 233–250
30. Ngo, C.-W., Jiang, Y.-G., Wei, X.-Y., Zhao, W., Liu, Y., Zhu, S., Wang, J., Chang, S.-F.: Vireo/dvmm at trecvid 2009: High-level feature extraction, automatic video search, and content-based copy detection. In: TREC Video Retrieval Evaluation Online Proceedings. (2009)
31. Jiang, Y.-G., Yanagawa, A., Chang, S.-F., Ngo, C.-W.: Cu-vireo374: Fusing columbia374 and vireo374 for large scale semantic concept detection. Columbia University ADVENT Technical Report #223-2008-1 (2008)
32. Chen, X., Mu, Y., Yan, S., Chua, T.-S.: Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In: Proceedings of ACM Multimedia. (2010)
33. Zheng, Y., Neo, S.Y., Chua, T.S., Tian, Q.: Probabilistic optimized ranking for multimedia semantic concept detection via rvm. In: Proceedings of CIVR. (2008)
34. Le, D.D., Poullot, S., Wu, X., Nett, M., Houle, M.E., Satoh, S.: National institute of informatics, japan at trecvid 2009. In: TREC Video Retrieval Evaluation Online Proceedings. (2009)
35. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. (2006)