TRECVID 2009 of MCG-ICT-CAS

- **Interactive search task: VideoMap: Interactive Search System by MCG-ICT-CAS**
  Juan Cao, Yong-Dong Zhang, Bai-Lan Feng, Lei Bao, Ling Pang and Jin-Tao Li
- **Copy detection task: TRECVID 2009 Content-Based Copy Detection by MCG-ICT-CAS**
  Yong-Dong Zhang, Ke Gao, Xiao Wu, Hon-Ttao Xie, Wei Zhang, Zhen-Dong Mao

VideoMap: Interactive Search System by MCG-ICT-CAS*

Juan Cao, Yong-Dong Zhang, Bai-Lan Feng, Lei Bao, Ling Pang and Jin-Tao Li

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{caojuan, zhyd, fengbailan, baolei, pangling, jtli }@ict.ac.cn

ABSTRACT

This paper describes the highlights of our interactive search system VideoMap for TRECVID 2009. To enhance the efficiency, the system has a map based displaying interface, which gives the user a global view about the similarity relationships among the whole video collection, and provides an active annotating manner to quickly localize the potential positive samples. Meanwhile, the system has powerful multiple modality feedback strategies, including the visual-based feedback, concept-based feedback and community-based feedback. These feedback algorithms can flexibly transform between each other by the automatic optimizing strategy. Finally the multi-strategies feedback achieves the best performance with MAP of 0.186.

Keywords

Video map, Interactive retrieval, Concept-based retrieval, Video community

1. Introduction

Why the system names as VideoMap? From the interface showed in Figure 1, we can see that VideoMAP visualizes the whole video database in a MAP as the background, where each row is a video and each shot in the video is represented as a thumbnail of keyframe. Based on the users' feedback and learning algorithms, these videos and shots are dynamically arranged so that the potential positive videos and shots are adjacent and in the center of screen.

Figure 1 The interface of VideoMap

VideoMap includes three kinds of feedback strategies: visual-based feedback, concept-based feedback and community-based feedback. The system can flexibly transform among these different feedbacks. To verity the performance of these algorithms, we have designed ten runs for the interactive search task of TRECVID 2009, and got the promising results shown in Table 1.

* This work was supported by the National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60902090, 60873165, 60802028), Co-building Program of Beijing Municipal Education Commission.

Run1: the multi-strategies feedback run with auto-switching.

Run2: the multi-strategies feedback run with expert –switching.

Run3: community based run with dynamic correlation score within communities.

Run4: advanced result based on Run5 by active concept labeling.

Run5: concept-based feedback by Graph-Based Multi-Space Semantic Correlation Diffusion (GMSSCD) method.

Run6: community based run by fusing local community reputation score and dynamic intra-community correlation score.

Run7: community base run with local community reputation score.

Run8: concept-based feedback by Distribution Based Concept Selection (DBCS) method.

Run9: the visual-time feedback run.

Run10: the visual baseline based on Latent Dirichlet allocation.

Table 1. the performance of ten runs for interactive search

Run_ID	MAP
I_A_N_MCG-ICT-CAS_1	0.186
I_A_N _MCG-ICT-CAS_2	0.169
I_A_N _MCG-ICT-CAS_3	0.112
I_A_N _MCG-ICT-CAS_4	0.149
I_A_N _MCG-ICT-CAS_5	0.109
I_A_N _MCG-ICT-CAS_6	0.118
I_A_N _MCG-ICT-CAS_7	0.080
I_A_N _MCG-ICT-CAS_8	0.076
I_A_N _MCG-ICT-CAS_9	0.139
I_A_N _MCG-ICT-CAS_10	0.118

2. Visual-based Feedback

Last year, we proposed a visual baseline in automatic search task [1] which achieves a satisfactory mean infAP of 0.028. This year we propose two runs based on visual features. The first one is the run10—visual baseline, which is also based on Latent Dirichlet allocation [2] following the last year's work. The second one is the run9—visual-time run, which fuses the temporal and visual similarity together.

2.1.1 The Visual baseline

In this method, we firstly represent each keyframe as a 500-dimensional soft-visual-keywords vector, which is from the work of City University of Hong Kong [3]. Then, we train the LDA to reduce the dimension of this feature to 80, which space is more semantic and more robust than the original feature space. Thirdly, the cosine similarity matrix between the keyframes is calculated offline for the interactive search. Finally, in the procedures of interactive search, we re-rank the shots based on their similarities to the labeled relevance shots and the query-examples to get the final visual baseline result.

2.1.2 The Visual-Time Run

As the neighboring shots seem to be semantic coherence, we combine the temporal and visual similarities together in run9, where the time decay window is used to measure the temporal similarity between shots in one video. As shown in the result of table 1, this run is much better than the visual baseline run. The results show that the temporal information is very important for video retrieval. Moreover, This is the main difference between video retrieval and image retrieval, which encourages us to find and use more essential characteristics of videos, not only simply represent the video as isolate keyframes.

3. Concept-based Feedback

Concept-based video retrieval has captured extensive research attention, mainly for its promising role in bridging the semantic gap. Different from the traditional content-based retrieval, concept-based video retrieval is enabled by pooling a set of concept detectors to bridge the gap between user semantics and low-level features, and is regarded as a kind of semantic content-based retrieval.

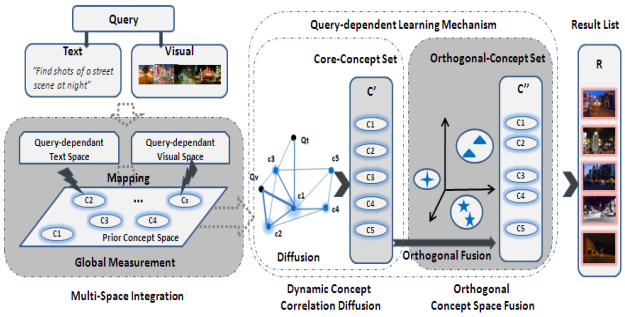

Figure 2. The framework of GMSSCD feedback algorithm

This year we also use the concept detectors results of CU-VIREO374 [3], and employ a distribution based concept selection method, named DBCS [4] as the baseline of concept-based method. Besides, we propose a novel concept-based method, named graph-based multi-space semantic correlation diffusion (GMSSCD), to explore the relationship between query and concept for video retrieval. As shown in Fig.2, given a multi-modal query, firstly we obtain a net-style relational model description by analyzing the organizational relations between query and concept and construct a concept correlation graph based on the net-style relational model. Secondly, a multi-mode mapping construction between query and concept is presented to enhance the pertinence of concept expanding. Thirdly, RoM (Ranking on Manifolds) [5] is extended to dynamically

expand index concepts, and finally the orthogonal concept fusion is used to retrieve video when a global stable state is achieved.

In addition, we also propose an active concept labeling strategy and design a flexible concept retrieval user interface (UI) for GMSSCD module. As shown in Fig.4, this UI allows users to directly label relevant and irrelevant concepts based on the labeled shots, and actively take part in the process of concept diffusion.

Figure 4 Interactive Concept Retrieval User Interface

To make a deeply analyses about the effectiveness of the GMSSCD, we design three concept-based runs described in section 1 to take comparison. From the official evaluation result in table 1, we can conclude that the effective retrieval algorithm and the efficient UI can help to enhance the retrieval performance, and also provide different users various flexible choices.

4. Community-based Feedback

One of the interesting properties of the WWW network and other types of networks such as collaboration and biological networks is the community structure [7,8,9], a structure of subgraphs with dense connections. In our perspective, communities also emerge in the visual correlation network of video shots from large collections because of the common semantic topics and visual contents shared between multiple videos. Discovered community structure as the tightly interrelated cluster of video shots with latent consistent semantics can be used to facilitate users in video search.

4.1.1 Community-based Active Annotation

In most existing interactive retrieval systems, only the shots recommended by the system are leaving for users to annotate. However, in our system the situation is quite different. As we known, the adjacent shots in MAP are similar and very likely to have the same label. Since that, the distribution of labeled samples in MAP can gives the user a global view about the similarity relationships among the whole video collection, and provides an active annotating manner to quickly locate the potential relevant samples region and actively annotate them, as illustrated in Figure 5. It is named Active Annotation.

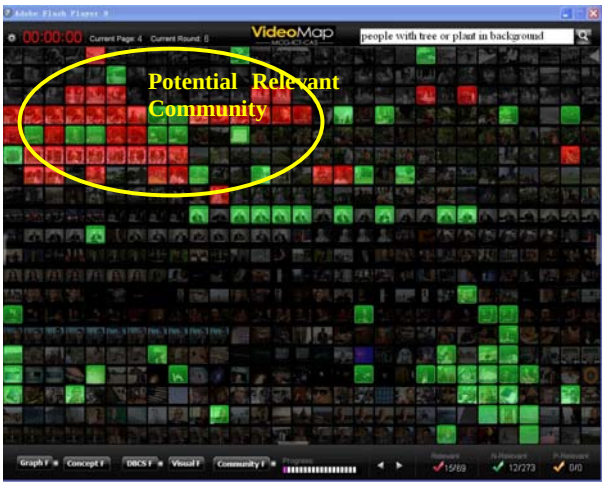

Figure 5. The User Interface of Active Annotation

4.1.2 Community-based Feedback

Besides the community-based interface, we also design three community-based feedback algorithms. First of all, we organize the video shots in a correlation network where vertices represent the video shots, and edges are weighted with pair-wise visual similarities computed in the above visual feedback method. Then, a random walk based community discovery algorithm [6] is performed on the correlation network to get the latent community structure where vertices have more influence to each other within the community than outside the community. At last, we design three community based runs with different ranking strategies to use the contextual community information and user feedback. We design three community-based runs to take a comparison. The framework is illustrated in figure 6:

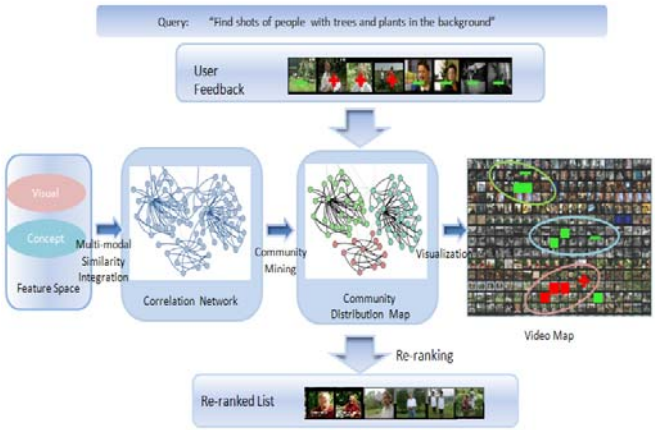

Figure 6. The framework of community feedback algorithm

In the run7, we give each community a reputation score according to the user feedback, and guarantee that community with more positive samples and less negative samples have a higher rank. Then video shots in the initial text based rank list are re-ranked by the reputation score of the local community it belongs to.

In the run 3, we focus on the dynamic similarity correlation of video shots within each community and update the score of video shots dynamically using the propagation result of a random walk process from the positive members in their corresponding community.

In the run 6, we fuse the local community score in run 7 and the dynamic correlation score in run 3 to get a complete measure of relevant video shots.

From the experimental results, we can see that dynamic correlation score is more effective than local community score for video search, and run6 outperforms the other 2 runs for it not only considers the static community structure in the collection, but also the dynamic relationship of video shots within the community.

5. Multi-strategies Feedback

As stated above, we explored several feedback algorithms, such as visual runs, concept-based runs and community-based run. Actually, every feedback algorithms have their own favorite scenario respectively. It is very important to choose an appropriate feedback strategy in proper time point, which will make the most of users' effort to improve the system performance. Since that, we submitted two runs with the multiple feedback strategies. The first one is the run2 -- expert-switching, where the system user is familiar with all the proposed feedback strategies. The second one is run1 – auto-switching, where the system automatically switches between different strategies by the suggestion of an automatic feedback strategy recommender. It is a good choice for the naïve users. The recommender is based on the current factors, such as: (1) the percentage of relevant samples in last pages; (2) the quality of the current strategy; (3) the percentage of relevant samples in all labeled samples; (4) the current feedback strategy; to provide the naïve user a wise suggestion.

The results in table 1 show that the multi-strategies feedbacks achieve best performance. Moreover, the run1 is some better than the run2, which demonstrates the effect of the automatic feedback strategy recommender.

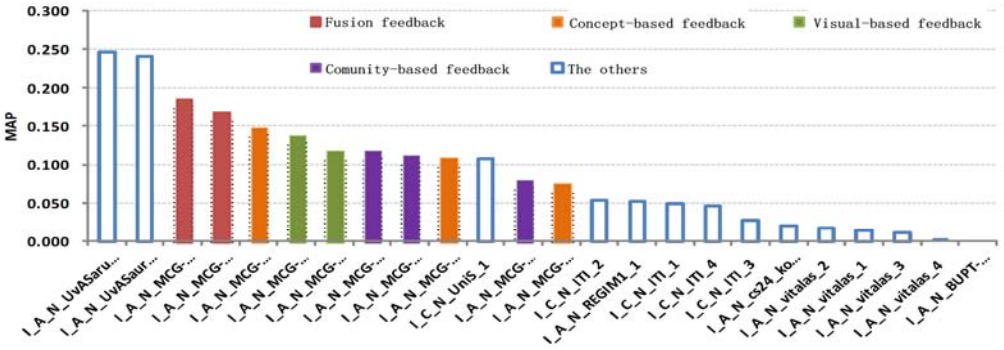

Figure 7 The performances of ten submitted runs for interactive search.

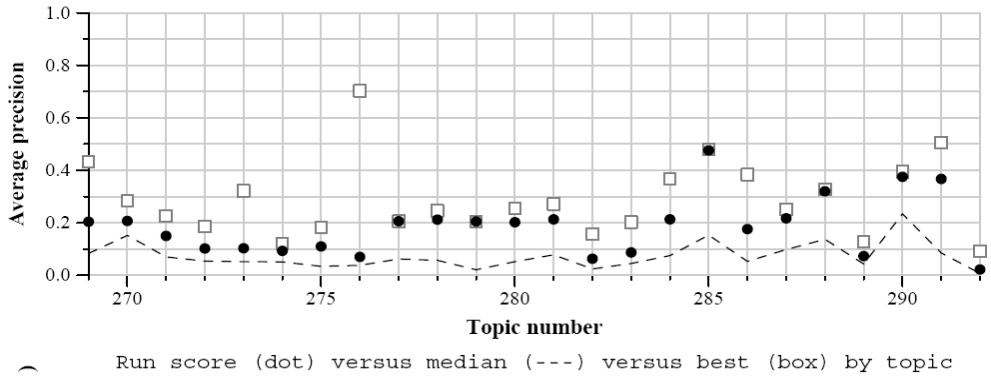

Figure 8 The performance for 24 topics of automatic multi- strategies feedback (run1)

6. Experiments and Analysis

We submitted 10 type A runs for interactive search task. Figure 7 shows their performances in all the runs[10], where the shaded bars are our submitted runs.

Firstly, we find that the multi-strategies feedback(run1 and run2) outperform the other proposed runs, in which only one strategy is used in the procedures of interaction. This means that switching to appropriate feedback strategy in proper time point actually improve the system performance significantly. Moreover, the figure 8 shows the detail results of the automatic multi-strategies feedback over 24 topics. It achieves nearly the best performance on six topics.

Secondly, we should notice that the three type of feedback strategies described above are complementary. The concept-based methods are popular as the concept detector bridged the "semantic gap". However, due to the limitation of the concepts size and the detector performance, these methods sometimes fail in some situation and the visual-based methods should be more effective. What's more, the community-based method discovers the community of the whole corpus and utilizes the distribution of labeled samples and relationship between shots. Since the information is not be used in the other two methods, it is an ideal complementary method. The complementary of these methods maybe the main reason that run1 and run2 work so well.

Reference

[1] J. Cao, Y.D. Zhang, B.L. Feng, X.F Hua, L. Bao, and X. Zhang, MCG-ICT-CAS TRECVID2008 search task report, Proceedings of TRECVID, 2008

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, January 2003

[3] Y.G. Jiang, C.W. Ngo, J. Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. ACM International Conference on Image and Video Retrieval (CIVR), Amsterdam, 494-501 (2007).

[4] J. Cao, H.F. Jing, C.W. Ngo, Y.D. Zhang. Distribution-based Concept Selection for Concept-based Video Retrieval. ACM International Conference on Multimedia (MM), Beijing, 19-24 (2009)

[5] D.Zhou, J.Weston, A.Gretton, O.Bousquet, B.Scho. Ranking on Data Manifolds. In Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS), Cambridge, MA, 169-176 (2003)

[6] P.Pons and M. Latapy. Computing communities in large networks using random walks. In ISCIS 2005, pages 284-293, 2005

[7] G.Flake, S.Lawrence, C.Giles. Efficient identification of web communities. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 150-160, 2000

[8] M.Girvan and M.Newman. Community structure in social and biological networks. Proc.Natl.Acad. Sci.USA 99:7821-7826,2002

[9] M.Newman. Scientific collaboration networks.i.network construction and fundamental results. Phys Rev E Stat Nonlin Soft Matter Phys,64(1 Pt2), 2001.

[10] A. F. Smeaton, P. Over and W. Kraaij, Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval(MIR '06), 2006

Content-Based Copy Detection by MCG-ICT-CAS*

Yong-Dong Zhang, Ke Gao, Xiao Wu, Hon-Ttao Xie, Wei Zhang, Zhen-Dong Mao

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{zhyd, kegao,wuxiao, xiehongtao, zhangwei, maozhendong}@ict.ac.cn

ABSTRACT

We participated in the content-based copy detection task in TRECVID 2009. This paper describes the details of our system for the task. We propose a series of measures to improve the precision, recall and detection speed of our system. To copy with various distortions, we integrate four modules which cope with one typical copy transformation respectively. To ensure the robustness and distinction of visual feature, we use the Block Gradient Histogram as frame descriptor, and utilize local descriptors with spatial context information to refine the primary result. Based on the above processing, we alignment frame sequence by checking Time Sequence Consistency to improve the precision of copy localization. Moreover, we present a novel GPU (Graphics Processing Units) based local interest point detector coined Harris-Hessian (H-H) to speed up local feature detection. Experiments show that these methods can enhance the accuracy and efficiency of detection, and make copy detection system adapt to real-time application.

Keywords

Content-based copy detection, Block Gradient Histogram, spatial neighborhood feature, hierarchical fusion method, sequence alignment, GPU based local feature detection.

1. System overview

To meet the need of practical application, we propose a series of methods to improve the precision, recall and detection speed of our system. The main contributions are shown as Figure 1.

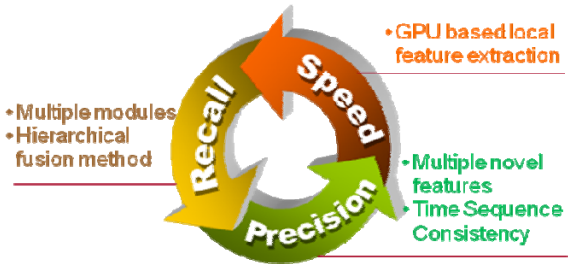

Figure 1: Main contributions of our CBCD system

Considering the transformations often used for generating the video copies [1] [2], we divide all the transformations into 4 types based on their primary transformations, and design 4 modules to cope with them respectively:

(1) Global quality decrease such as blur, adding noise, change of gamma、resolution and contrast, etc.

(2) Partial content alteration such as occlusion, shift, crop, and insertions of pattern (including the Picture in Picture type 2, the original video is in the background).

(3) Picture in picture type 1(The original video is inserted in front of a background video).

(4) Flip (vertical mirroring).

The quality of copy detection not only depends on the type of transformations, but also on the property of query segments. Because most of the query segments are very short, and do not have much motion and activity, the motion information are not used in our system. A candidate video sequence is defined as a set of successive keyframes described by features (global features or local features). We use these features to progress an approximate search [3] in the database and get the similarity of keyframe-pairs. Afterwards, the Time Sequence Consistency method is used to locate the copies' boundaries and find their temporal position. At last, the detection result of each module is fused to make the final decision. Moreover, to speed up local feature detection which is the most time-consuming, we present a novel GPU based local interest point detector called Harris-Hessian (H-H). The flowchart of our system is shown as Figure 2.

* This work was supported by the National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60902090, 60873165, 60802028), Co-building Program of Beijing Municipal Education Commission.

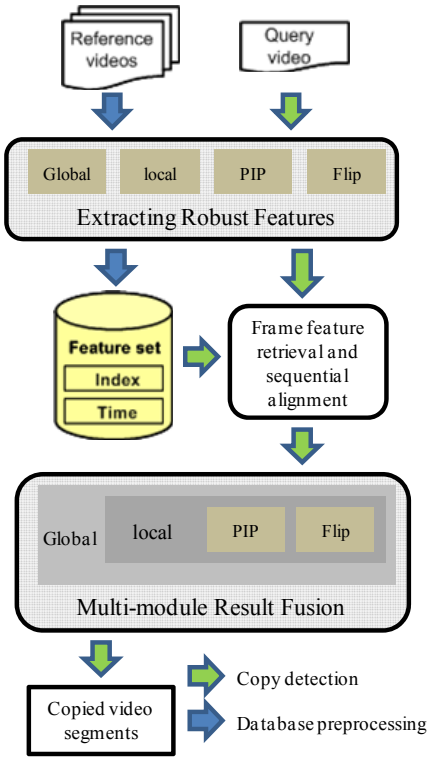

Figure 2: Flowchart of our CBCD system

2. Multiple Novel Features

In order to copy with various distortions, we have proposed multiple kinds of features. To ensure the robustness for global transformation (e.g. blur, noise and color changing), our system uses the Block Gradient Histogram as frame descriptor. For various local image distortions (e.g. partial occlusion, crop and shift), we utilize local descriptors with spatial context information to refine the primary detection result. Besides, some special processes are also proposed to deal with transformations such as PIP and Flip.

2.1 Block Gradient Histogram

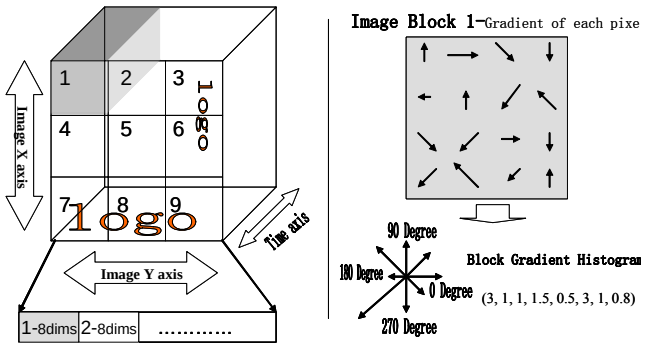

Figure 3: Illustration of Block Gradient Histogram

We extract Block Gradient Histogram features form each key frame of video shots. Motivation of this is to rapidly search candidate copies which are similar to query video in global frame appearance. Here we extract DC coefficients of intra frames from MPG1 video, and divide each keyframe into 9 blocks. Then global gradient histogram of frames (8*9=72 dimensions) is extracted as visual feature (ref. to Figure 3). The procedure need not decode the video and greatly accelerates the processing. We also eliminate letter box in frames to improve the feature discrimination power. The block gradient histogram based feature is invariant to certain transformations, e.g. blur, noise, lighting, color and global changes. Using more frames and conducting frame-to-frame matching could guarantee the accuracy of copy location.

2.2 Local Features with Spatial Information

To deal with partial content alteration such as occlusion and crop, we detect salient local point and extract 8-dims gradient histogram from each local patch (similar with the right part of Figure 3).

In order to reduce the illegibility of local feature, we present a method to describe its spatial information, as in Fig 4. First, the neighborhood of each local point is divided into 4 blocks (scale is 2 times of the local patch), and these blocks are sorted using their average gray level, something resembling [4]. The rank of each block is used as the spatial signature of each local point, which is compact and distinguishing. With the spatial constrain, we can effectively reduce most false matching of local features.

Second, given a query keyframe with local features $\{q_1, q_2\ldots, q_n\}$, we will obtain several candidate matching keyframes KF_i denoted by $\{C_{i1}, C_{i2}\ldots, C_{in}\}$. If we divide the bounding box containing these local features into N*N blocks, then the spatial distribution histogram can be easily calculated as Figure 4. With this information, we can further reduce the false matching of local features.

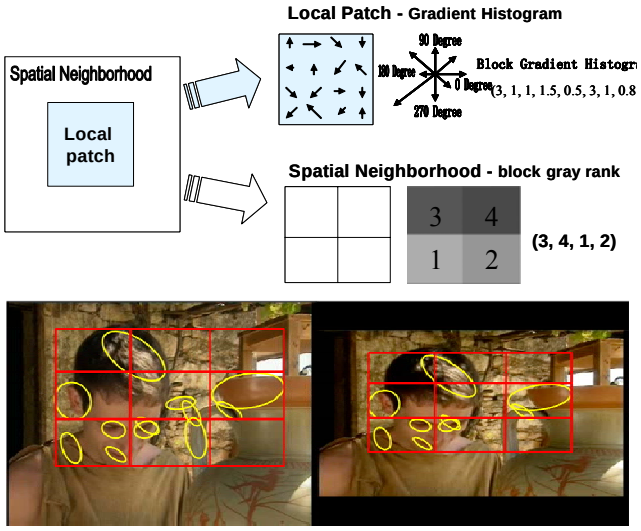

Figure 4: Local feature with spatial information (Upper: for each single feature. Below: for a set of matching features)

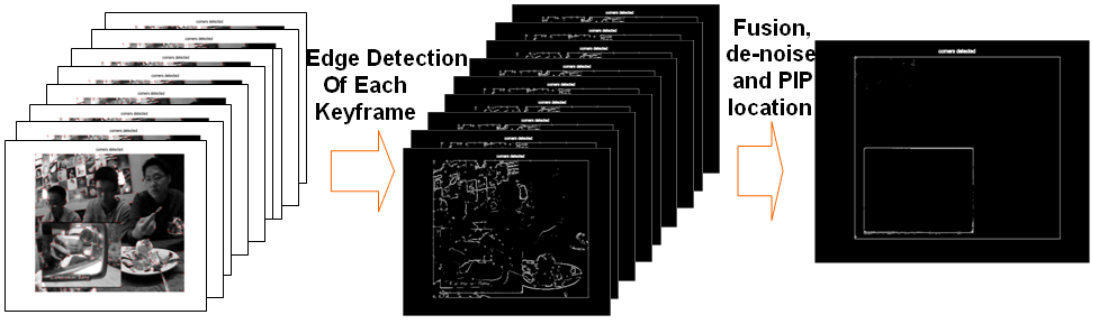

Figure 5: Illustration of PIP boundary location.

2.3 Feature for PIP

Picture in picture type1 (PIP) is the original video been inserted in a front of a background. As the smaller video only takes little part of the whole image, i.e. only 20 percent of the original image, common methods can't perform well under this situation. Therefore, we design a module for this kind of transformation separately: adaptively locate the PIP boundary based on edge detection and fusion method, as shown in Figure 5. After that, Block Gradient Histogram method for Global transformation is used for PIP copy location.

Firstly, we get the key-frames based on shot boundary detection. Secondly, multi-scale and adaptive canny algorithm [5] [6] is used to detect edges for each key-frame. Thirdly, we fuse these results to get a fusion boundary of PIP. Fourthly, with Probabilistic Hough Line Detection, we get all the potential line segments for the PIP region. Finally, we use rule-based method to determine the PIP borders. After this, the detection result is returned.

2.4 Feature for Flip

As to the frame flipping change (i.e. vertical mirroring of frames), we use a very simple method to deal with it. Because there is only difference in left and right position, so we only change some dimensions of the local and global feature to simulate flipping, and the other work is almost the same. That is to say, for some candidate queries which cannot be determined as a copy based on original features will be matched further based on these flipping features.

3. Time Sequence Consistency method for Optimal Copy Location

Given a query video, the features form its keyframes are extracted and matched with dataset. Considering the detection accuracy and speed, we first use ANN indexing structure to find the similar keyframes. And then, the time sequence consistency will be used to refine the copy location.

3.1 Feature Indexing Structure

The ANN indexing structure is adopted in our framework for feature storage and search. First ANN algorithm builds database to store the large scale frame corpus. Then we use rapid Q-Range Nearest Neighbor search (NRNN) to query the database, which returns all reference frame features in feature space whose distance to the query is smaller than Q. The expected time for a search is logarithmic in the number of elements store in the database.

3.2 Time Sequence Consistency method

Even with various features mentioned above, there are still too many similar keyframes which would result in false matching and copy location error. Therefore, we propose a time sequence consistency method for optimal copy location. Experimental results have shown that it is one of the curial elements for performance improvement.

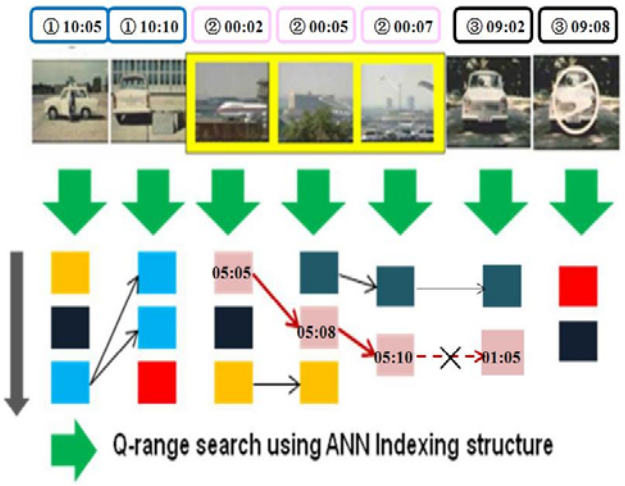

Figure 6: Time Sequence Consistency method

As shown in Figure 6, by searching database using n query frame feature, n result lists are obtained and each list contains L_i (i=0, …, α) reference frames. α is fixed to be 500 to limit computation cost. Using these results as nodes and the feature similarity as node weight, we construct a graph for copy location. Edges are added if temporal constraints are satisfied: 1) frames f_a and f_b belong to the same reference video; 2) time stamp difference $0 <= T_b - T_a <= \mu$. μ is empirically set as 10 (sec). A copy will be located if total weight of the maximum flow in graph is larger than β (fixed to 20 empirically) as in formula (1) (2).

$$M = \max_{i,j,X} \sum_{l=i}^{j} weight(node_l) \quad (1)$$

$$location = \begin{cases} frame_i \ \ldots \ frame_j \ \ of \ \ videoX, \ \ if \ \ M > \beta \\ none \ \ \ \ else \end{cases} \quad (2)$$

4. Result Fusion

We have tried number of fusion method including non-hierarchical method and hierarchical method, as shown in Figure 7 (the upper one is non-hierarchical method, and the one below is hierarchical method). Non-hierarchical method means we use 4 modules to calculate each query separately at the same time, and only the result with the maximal score will be submitted. It is very simple but not efficient enough, because the global feature based method is very fast and effective, so we can use it to filter some result. Accordingly, the hierarchical method submits the result of each module in turn. For each query video, if any previous module has found its corresponding video clip, then we will submit the result, and go on to calculate the next query. Tested in the develop set, hierarchical method performs the best for most test queries, and the processing time is reduced greatly.

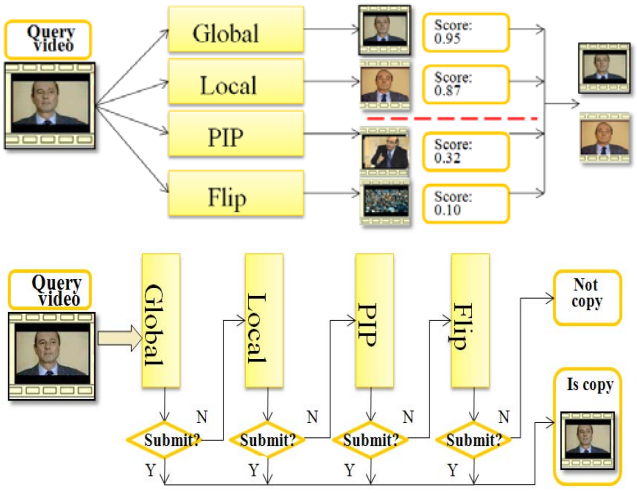

Figure 7: Illustration of result fusion methods.

5. GPU based Local Feature Extraction

It is well known that local feature extraction is very time-consuming due to the limited computing ability of CPU. To take full advantage of the powerful computing capability of graphics processing units (GPU) to speed up local feature extraction, we present a novel GPU based scale invariant interest point detector, coined Harris -Hessian(H-H). H-H detects Harris points in low scale-space and refines the location and the scale in higher multi-scale space with Hessian matrix. Compared to the existing methods, H-H significantly reduces the computation complexity and has better parallelism. The experiment results show that with the assistance of GPU, H-H achieves up to a 10-20x speedup than CPU based method. It only takes 6.3ms to detect a 640*480 image with high detection accuracy, meeting the need of real-time detection.

Although a few GPU-based local feature extraction methods have been proposed [7-9], they just implemented the original algorithms on GPU. As the original methods are complex and need a lot of data copy between different memory, which is not suited to the architecture of GPU and may become a performance bottleneck in CUDA. In this paper, we propose a new scale invariant interest point detection algorithm according to the characteristics of GPU and CUDA. The new algorithm detects interest point with Harris detector and refines the point relying on the determinant of Hessian matrix. It has good parallelism and can satisfy the demand of real-time detection under large scale data with high detection accuracy.

5.1 Harris -Hessian Method

We first detect Harris points in low scale space σ_0 to get the candidate point set C_i. σ_0 can be equal to 0, which means detecting corners in the original image. σ_0 may choose value in a small range of sale.

After obtaining C_i, for each point we apply the DET value of the Hessian matrix to remove the false detection and confirm the scale of interest points simultaneous. Given a point $X = (x, y)$ in an image I, the Hessian matrix $H(X, \sigma)$ in X at scale σ is defined as follows:

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix}, \quad (1)$$

where $L_{xx}(X,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point X, and similarity for $L_{yy}(X,\sigma)$ and $L_{xy}(X,\sigma)$. The DET value is:

$$DET(H) = \left| L_{xx} * L_{yy} - (0.9 L_{xy})^2 \right| \quad (2)$$

The steps of false detection elimination and scale confirmation are as follows:

1. For each candidate point m in C_i, calculate its $DETs$ under a range of scale σ_n.

   $\sigma_n = 1.2^n \sigma_0 (n = 1,2,\ldots,N), \sigma_0 = 1.2$. The value of N is determined by the specific applications.

2. If the $DETs$ of m attain no extremum in a series of σ_n, we reject it.
3. If m has maximum DET value in scale σ_n and that DET is greater than a threshold, then m is an interest point and its scale is σ_n.

$$DET(X,\sigma_n) > DET(X,\sigma_{n-1}) \wedge DET(X,\sigma_n) > DET(X,\sigma_{n+1})$$

$$DET(X,\sigma_n) > T \quad (3)$$

Compared to existing methods, the algorithm designed in this paper has the following advantages:

- ✧ Computational complexity is reduced significantly.

  Instead calculating each pixel's response, we only focus on a few candidate points detected in low scale and the computational complexity is $O(n)$.

- ✧ Our method has better parallelism and is more suitable for GPU implementation.

  Our method only needs a few times of kernel invocation and data transfer. So it makes full use of shared memory, resulting in good performance.

5.2 Implementation on GPU

As H-H method only involves convolution and differential operation, we partition the source image equally into data-blocks and distribute them among the thread blocks, as illustrated in Fig.8. In our realization, there are 256 threads in a thread block, so the total number of thread blocks is:

$$Block\_num = \frac{image\_width}{16} * \frac{image\_height}{16} \quad (4)$$

To speed up the process of the convolution, we approximate the Gaussian with box filters, which is beneficial to GPU acceleration too. As box filters are constant, we put them in the constant memory when the program starts to reduce the frequency of data transfer.

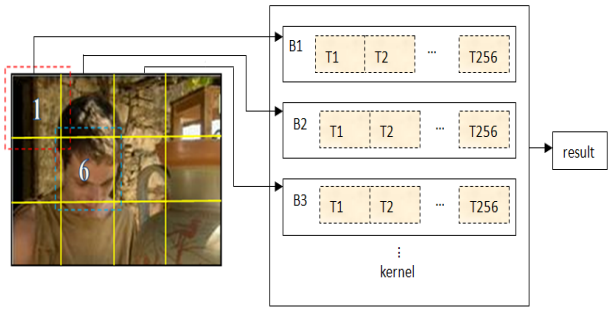

Figure 8: Illustration of data partition using GPU

During the calculation of convolution, the problem of "boundary cases" will be confronted. As shown in Fig.2, when calculating convolution for blocks 1 and 6, the required data are represented by the red and blue dashed box. So we employ texture memory, which handles the "boundary cases" automatically.

The execution flow of H-H in GPU is shown in Fig.9. Firstly, the data of source image and box filters are copied into texture memory and constant memory; then each thread gets its data and starts executing; the intermediate results are stored in shared memory, through which the threads in a thread block coordinate with each other; and then all threads continue executing until the task is completed; at last, the result is transferred to the DRAM.

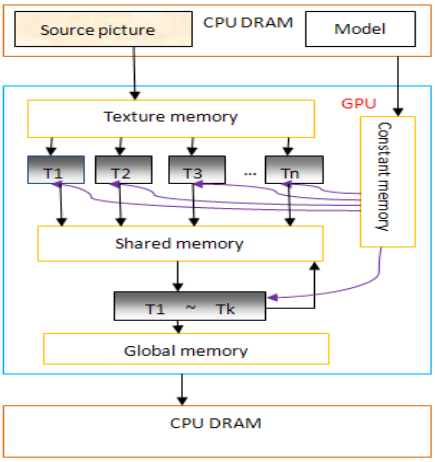

Figure 9: Execution flow of H-H method using GPU

5.3 Performance on GPU

The experiment environment is: Intel Core E8400 3.0GHz with 2048MB memory, NVIDIA GeForce 9800GTX+ with 512MB DRAM, Microsoft Windows XP sp2, CUDA Toolkit 2.1 and CUDA Driver (181.20).

Our dataset includes 11250 images, which include nature, people, news, video, drawing and cartoon. Their size ranges from 256*364 to 1024*1024. To construct test images, we add attack such as crop, contrast change, occlusion and zoom to them randomly.

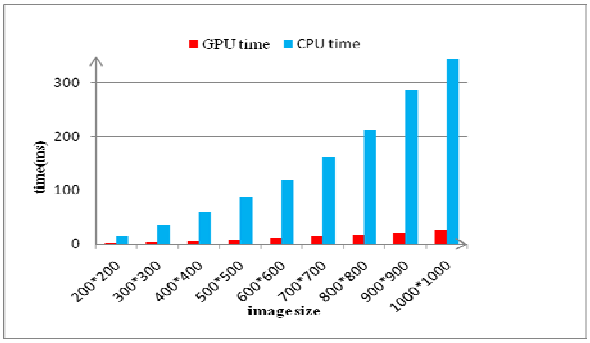

Figure 10: Time cost for CPU and GPU to detect H-H point

Fig.10 shows the time cost of GPU and CPU to detect the H-H points in images of different size. Compared to CPU, the GPU realization achieves up to a 10-20x speedup. It only takes 6.3ms to detect a 640*480 image, getting 437points. It demonstrates the good parallelism of H-H.

6. Copy Detection Experiments and Analysis

We submitted a total of 3 runs. The difference of them lies in the fusion methods and parameters. The result of our best run (hierarchical method ICTBAOPT) and the detection speed are shown as Fig 11. We can find that our methods can enhance the detection accuracy and efficiency obviously, and the introduction of GPU based local feature extraction is very necessary.

References

[1] H. Jegou, M. Douze, and C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in ECCV, 2008.

[2] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, Video copy detection: a comparative study, in CIVR, 2007.

[3] K.I.Lin and C. Yang, The ANN-Tree: An Index for Efficient Approximate Nearest-Neighbor Search, In Conf. on Database Systems for Advanced Applications, 2001.

[4] C. Kim and B. Vasudev, Spatiotemporal sequence matching for efficient video copy detection, IEEE Trans on CSVT, 2005.

[5] C. Harris, M. Stephens, A combined corner and edge detector. proceedings of fourth alvey vision conference. 1988, 147-151.

[6] J.Canny, A computational approach to edge detection, IEEE Transactions on Patten Analysis and Machine Intelligence, 1986, PAMI-8(6):679-698.

[7] S. Heymann, et al. Sift Implementation and Optimization for General-Purpose GPU, Proceedings of the 15th WSCG, 2007

[8] S. Sinha, J. Frahm. GPU-based Video Feature Tracking and Matching, Workshop on EDGE, 2006.

[9] N. Cornelis and L.Van Gool, Fast scale invariant feature detection and matching on programmable graphics hardware, CVPR Workshop, June 2008.

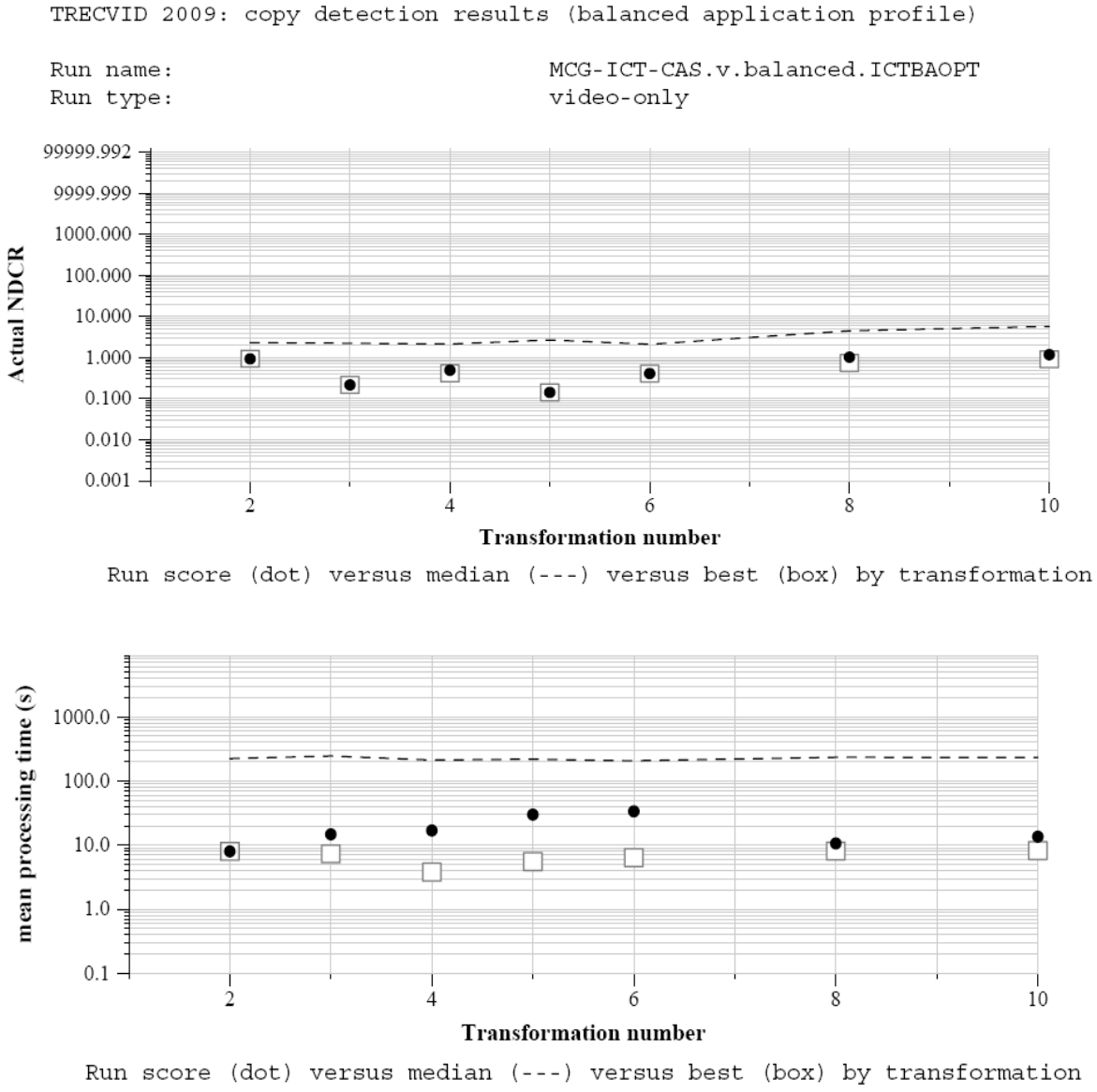

Figure 11. Performance of our system in TRECVID2009_CBCD.