Place Retrieval with Graph-based Place-View Model

Xiaoshuai Sun, Rongrong Ji, Hongxun Yao, Pengfei Xu, Tianqiang Liu, Xianming Liu
Visual Intelligence Lab, Harbin Institute of Technology
No.92, West Dazhi Street, Harbin, P. R. China
86-451-86416485, P.O.BOX 321, Harbin Institute of Technology, 150001
{xssun, rrji, yhx, pfxu, tqliu, xmliu}@vilab.hit.edu.cn

ABSTRACT
Places in movies and sitcoms could indicate higher-level semantic cues about the story scenarios and actor relations. This paper presents a novel unsupervised framework for efficient place retrieval in movies and sitcoms. We leverage face detection to filter out close-up frames from video dataset, and adopt saliency map analysis to partition background places from foreground actions. Consequently, we extract pyramid-based spatial-encoding correlogram from shot key frames for robust place representation. For effectively describing variant place appearances, we cluster key frames and model inter-cluster belonging of identical place by inside-shot association. Then hierarchical normalized cut is utilized over the association graph to differentiate physical places within videos and gain their multi-view representation as a tree structure. For efficient place matching in large-scale database, inversed indexing is applied onto the hierarchical graph structure, based on which approximate nearest neighbor search is proposed to largely accelerate search process. Experimental results on over 36-hour *Friends* sitcom database demonstrate the effectiveness, efficiency, and semantic revealing ability of our framework.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing, H.3.3 Information Search and Retrieval.

General Terms
Algorithms, Design, Experiment

Keywords
Video Analysis, Place Retrieval, Face Detection, Saliency Map, Key Frame Clustering, Graph Partition, Inversed Indexing, Large-Scale Retrieval

1. INTRODUCTION
The explosive growth of internet technology in the past decade has witnessed the constitution evolution of web information from solely text to rich multimedia, including image, audio, video and their combinations. Among them, video data consists of a large portion of web multimedia and still enjoys promising extension speed. With the prevalence of online video sharing websites such as YouTube [1] and Blinkx [2], there are increasing gigantic volumes of videos on the web. For instance, Blinkx [2] declares owning over 26 million hours of videos that can be displayed online. However, most of these video resources are not well labeled by textual descriptors that reflect their content information to facilitate retrieval. On one hand, the gigantic volume of video data available on the web restricts the application of manually labeling. On the other hand, the semantic gap [3] between user understanding and video content representation largely restricts the application of automatic computerized labeling. The incomplete labeling of video data calls for efficient content-based search technique to facilitate user access and browsing within video database.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-312-9/08/10...$5.00.

Near-duplicated video retrieval aims to search identical or almost identical video clips based on their visual contents. In such scenario, video shot is usually treated as the basic element in similarity matching. Generally speaking, video shots usually consist of several foreground actions and static/moving backgrounds within a consecutive camera motion. However, it is not always reasonable to adopt solely shot-based matching to search user targets. Especially, users are usually with specific search purpose: They may be interesting in searching identical actor, motion, or background place. This is extremely true for the case of movies and soap operas, in which the scenarios are usually happened only for certain actors, within certain places. Figure 1 shows an example, in which places such as "*Central Perk*" *coffee house*, *Monica's apartment*, and *Joey's apartment* are frequently appeared in *Friends* soap opera and contains several views that physical connective to constitute this place, each of which owns almost identical appearance over the 10 seasons of this soap opera. Viewers in many cases would rather prefer to browse the stories that happen within the identical place, rather than actors or motions.

Figure 1. Scenarios in soap opera *Friends* usually conduct in certain places. Each place has multiple views or scenes that are physically connective and usually shot concurrent. Top raw: The Monica's apartment; Middle raw: The "Central Perk" coffee house; Bottom row: The Joey's apartment.

To the best of our knowledge, place-level video retrieval is out of research concern in former literatures of video search research. Referring to place retrieval, our method is different from methods target on scene classification and recognition [4, 5], in which scene categories are pre-defined. Siagain [6] leverage saliency map from

image to represent predefined scene and adopt PCA to distill representative features from the training set of each scene, based on which supervised classifier is learned in scene classification. Bag-of-Region representation is leverage in scene classification [7], which is demonstrated to be effective in the case of partial occlusion and appearance variance. Normalized Cut is adopted in [8] for graph-based scene grouping. Our method differs from all above-mentioned works: On one hand, works in [4-8] aim to *classify* scenes into several pre-defined categories and usually involves training parses. In contrast, our work aims to *rank* with no pre-defined categories in an unsupervised manner. On the other hand, place is a broader definition comparing with scene, for which one place may be includes several location-connective views (scenes) that are physical consecutive. This is indeed not well investigated in former papers [4-8].

In this paper, we discuss the issue of place-level video retrieval and present a unified framework to address the key challenges in this issue. To clarify before our discussion, the formalized definition of "Place" is given as follows:

Definition 1 (Place): Place in movie or sitcom is a physically consecutive space that frequently occurs within shots, owing to consecutive camera motion.

Definition 2 (Place Retrieval): Given a query shot by user, place retrieval aims to retrieve video shots that with identical or near-identical background place. The identical definition is from physically consecutive viewpoint, which means that the view appearance is not guaranteed to the same, but there location is conjunctive together. Indeed, the variant appearance but physically connectivity would reflect place semantic at a higher level.

Generally speaking, there are four challenging issues in place-level video retrieval:

1) Distillation of background place from foreground objects and actors, especially in the existence of global motion.

2) Appearance representation that is robust to scale and viewpoint variance. Especially, the spatial dependence of regions in view is importance cue in appearance modeling.

3) Multi-view representation and modeling of place appearances, and their corresponding retrieval strategy.

4) Search efficiency in large-scale video database, in which linear-scanning would become unrealistic for application.

This paper presents a novel place retrieval framework that enables efficient search of near-duplicated places from movies and sitcoms in large-scale scenario. To address the first issue, face detection is leverage to filter close-up shots from video database. Consequently, spatial & temporal saliency analysis is utilized to remove foreground motions from shot contents. To address the second issue, pyramid-based spatial-encoding color correlogram is presented for robust place appearance representation. By clustering views and model co-place relation by inside-shot associations, the multi-view issue of an identical place is addressed using graph-based normalized cut partition. Finally, we address the forth issue by inversed indexing of the hierarchical graph structure, based on which approximate nearest neighbor search is proposed for efficient retrieval in large-scale database. The flaw chart of our proposed place retrieval system is presented in Figure 2.

The rest of this paper is organized as follows: Section 2 presents our place extraction algorithm, including close-up filtering and foreground removal. Section 3 presents our view clustering and shot concurrent analysis algorithm for graph-based place generation, together with our novel graph-based hierarchical inversed indexing algorithm and approximate nearest neighbor search to facilitate large-scale application. Extensive experiments over 36-hour *Friends* soap opera are presented in Section 4. Finally, this paper concludes in Section 5 and discusses our future research directions.

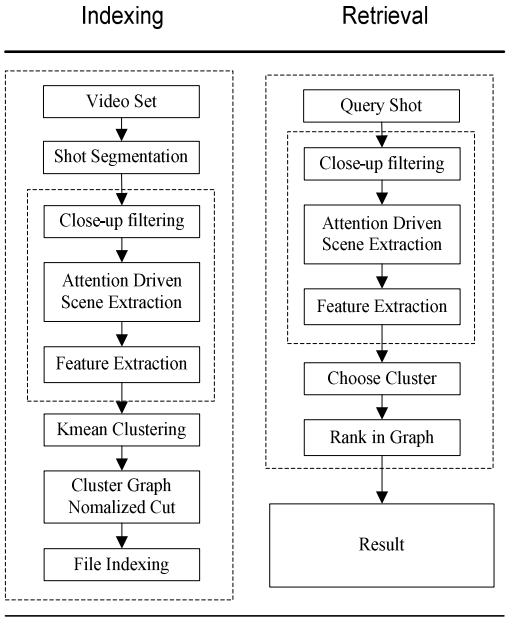

Figure 2. System flaw chart of our proposed framework.

2. PLACE DISTILLATION FROM MOVIES

The first step of our proposed framework is the distillation of place views from the video shots. In this step, the challenges lie in the moving foregrounds (maybe multiple) and viewpoint & scale variances. Our place distillation algorithm consists of two steps: Firstly, we filter close-up shots from our video dataset, which do not reflect sufficient place information for both modeling and retrieval. In this step, face detection [9] is adopted to locate actors from video shots, and the occupying ratio of facial region is calculated for filtering close-up shots. Secondly, spatial & temporal saliency analysis [10] is adopted to select and remove the foreground motion, based on which the background views are distilled for place generation. Figure 3 presents the flowchart of our proposed place distillation algorithm.

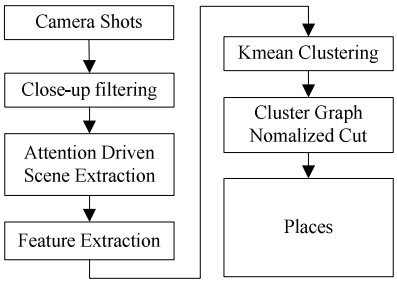

Figure 3. Flaw chart of place distillation from movies

2.1 Close-Up Frame Filtering

As the preliminary step, shot boundary detection (SBD) is achieved by leveraging graph partition model [12], which is demonstrated to be effective in TRECVID SBD competition.

Within each key frame of a shot, we leverage the face detection algorithm in [9] to detect the face regions that appear at this key frame. The occupying ratio of face region is calculated using Equation 1:

$$R_{occupy} = \frac{Area(Face)}{Area(Key\ Frame)} \quad (1)$$

Once this occupying ration R_{occupy} is larger than a given threshold T, a close-up frame is defined, which means that this frame basically consists of large portion of actor faces and bodies. Consequently, such frame only contains limited background information. Since these close-up frames usually occupied by large portion of static foreground regions, such as clothes and bodies, they would largely affect the consequent background generation process, not matter the background extraction algorithm using either GMM [13] or saliency map [11]. We filter all close-up frames out of our shot database to remove its negative affection.

2.2 Place Distillation using Saliency Analysis

Background modeling and extraction is a difficult problem in video analysis field. Literature in background modeling usually adopts Gaussian Mixture Modeling (GMM) [13] for foreground abstraction. However, to address the global motion, motion compensation should be adopted for static background recovery, which usually results in not stable background results. Another solution is the moving object tracking and target region erasing. However, this approach is still inconsistent to handle multi-object movements and target in/out. In this paper, different from above-mentioned strategies, we leverage saliency map analysis to achieve almost real-time place distillation, which also merits in its effectiveness in address the global motion problem.

The spatial-temporal saliency region is co-determined by fusing spatial-temporal FOA [10] to capture the attention focus as presented in Figure 4. For each frame of a detected shot, the spatial saliency map is constructed by saliency-based computational model in [11]. For the successive frames within this shot, the motion saliency map is constructed based on dynamic attention model [10]. The spatial-temporal saliency region is co-determined via the fusion of spatial-temporal FOA [10] to capture the attention focus.

Based on the former-constructed saliency maps, the salient foreground motion (usually is actor motions) is located and removed from each video frame: We define an erasing rectangle as in Equation 2 for foreground removal:

$$Rec_{X,Y} = \{\forall x_i \in X, \forall y_i \in Y \mid x_{min} \le x_i \le x_{max}, y_{min} \le y_i \le y_{max}\} \quad (2)$$

in which $Rec_{X,Y}$ represents the minimal bounding rectangle of the detected saliency region. Based on the frame that removes the first most salient region, we iteratively this removal process to erase the second most saliency region also using Equation 2. This process is iterated by our algorithm until the following stop criterion is satisfied:

$$T_{stop} = 1\ iff.\ \{\frac{Area(Remaining)}{Area(Key\ Frame)} < T_{Area} | S_{sailency} < T_{saliency}\} \quad (3)$$

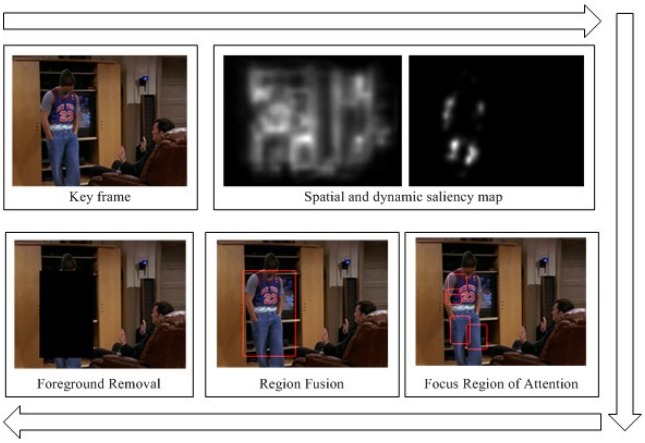

Figure 4. Attention-driven foreground removal.

In Equation 3, T_{stop} is the region removal stop threshold; *Area(Remaining)* is the area of remaining region that are possibly be background; *Area(Key Frame)* is the original area of key frame; T_{Area} is the area threshold; $S_{saliency}$ is the saliency strength of currently concerned region, $T_{saliency}$ is the saliency threshold. From the condition constraints of Equation 3, we could see that the first part is to restrict that there should not be too limited area in background region, which means the background is almost occluded by foregrounds. The second part of this constraint is the saliency restriction to maintain background in the case of global motion. Although our algorithm would results in partial background missing and partial foreground remaining, these two issues would be eliminated using a robust pyramid matching algorithm presented later in Section 3. Figure 5 shows a visualized example of our iterative saliency region removal algorithm.

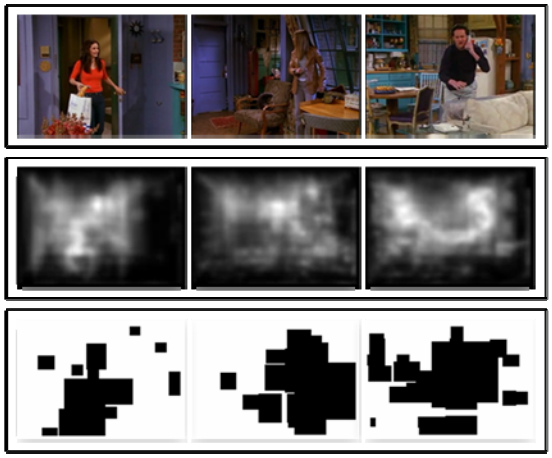

Figure 5. Visualized example about iterative saliency region removal for background place extraction. The first row is the original key frame, the second is spatial temporal saliency, the last one is the final background mask.

2.3 Place Description

Considering the movie and sitcom edition processes, there are two following specific characters in place appearance differ from traditional object search and recognition task:

1) The rotation variance is seldom appeared for places in movies or sitcoms. As a result, the spatial dependence and orders is much fixed in designing effective feature extraction algorithm.

2) Scaling and viewpoint are changeable in place appearance. This

results in the effectiveness loss in plan-level local feature matching [7, 8].

In this paper, we propose a novel place appearance extraction algorithm, named: **P**yramid-Based **S**patial-**E**ncoding **C**orrelogram (**PSEC**) in robustly place feature expression. Firstly, we hierarchical partition each frame into equal-sized regions, and linear scanning this region sequence (left to right, top to down) to encode their spatial dependency. Secondly, within each region of each hierarchical level, we extract an 8-bin color correlogram [16] within each region. The extraction process of pyramid-based spatial-encoding correlogram is presented in Figure 6.

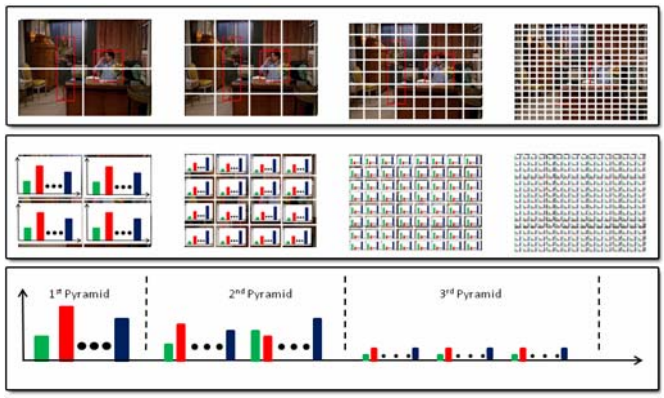

Figure 6. The extraction of pyramid-based spatial-encoding correlogram.

For PSEC-based similarity matching, multi-scale comparison between frame *i* and *j* is achieved using Equation 4 as follows:

$$Similarity(i,j) = \sum_{k=1}^{K} w_k \sum_{l=1}^{L} || H^{i}_{k,l} - H^{j}_{k,l} ||^2 \qquad (4)$$

in which, K is the total pyramid levels; w_k is the weight of k^{th} level matching; L is the bin number of each color correlogram; $H^{i}_{k,l}$ is the correlogram value in l^{th} bin, k^{th} level, i^{th} frame; $H^{j}_{k,l}$ is the correlogram value in l^{th} bin, k^{th} level, j^{th} frame. The spatial relationships of regions are encoded into order of correlogram bins at each level. Using pyramid-based color correlogram matching, the scale and viewpoint variance would largely degenerated.

3. GRAPH-BASED UNSUPERVISED PLACE APPEARANCE MODELING

This section describes our place appearance modeling and indexing algorithm to facilitate effectively place search while maintaining its efficiency in large-scale database. We challenge the task of near-duplicated view grouping and physically conjunctive view (of identical place) associating using a two step graph-based algorithm. The first step is the near-duplicated view clustering; the second step is the connective analysis using physically conjunctive information from camera motion connectivity within identical shot.

3.1 View Clustering

Based on the similarity metric as presented in Equation 4, we leverage k means clustering to group visually similar background views into clusters. In this step, we settle a large clustering number and control the frame number within each cluster by considering both frame number and clustering variance using Equation 5:

$$T = Num_{frames} + \lambda \times Variance_{cluster} \qquad (5)$$

in which Num_{frames} represents the number of frames within this cluster, $Variance_{cluster}$ is the frame variance within this cluster. λ is the combination factor that fuse the considerations of frame numbers and their variance within the cluster to decide the rejection threshold T, once larger than this given threshold, we stop adding photos into this cluster. Based on Equation 5 and using a large clustering number, our clustering algorithm gains a large amount of view clusters and almost near-duplicated views within each cluster.

3.2 Place Graph Generation by Shot Concurrence Analysis

Different views maybe comes from the same place with different viewpoints; the next step is to associate these place-identical views together. Recalling that each shot in movie or sitcom consists of consecutive camera motion over identical place, it offers us an effective cue in place-identical view conjunction.

Firstly, based on the cluster centers, we construct a graph representation <*V, L*> among views, in which *V* is the set of vertexes each represents a cluster, *L* is the set of links each represents a link between two vertexes. The algorithm of connective view discovery is presented as in Table 1:

Table 1: Algorithm for graph-based view connectivity discovery

1. **Input:** Graph <*V,L*>
2. **For** each remaining shot in the video dataset after close-up filtering and foreground removal**{**
3. Extract its key frames <*K*> using the algorithms as in [14].
4. **For** each key frame *k* with in <*K*>**{**
5. Find its belonging view cluster C_1
6. For other key frames within <*K*>**{**
7. Find its belonging view cluster C_2
8. Increase the connectivity value of link l_{12} by 1
9. **}**
10. **}**
11. **}**
12. Leverage normalized cut [15] to partition graph into groups, each of which is defined as a *Place*.
13. **Output:** Places grouping results, each of which has certain number of view clusters.

Especially, the normalized cut [15] algorithm is adopted to separate the graph nodes into groups, each of which is defined as a *Place*. Furthermore, we structuralized this graph-based place-to-view partition, and reversed index each shot into the view node in this structure. It results in a hierarchical inversed indexing model, which we name as **G**raph-based **H**ierarchical **P**lace-**V**iew (GHPV) model.

As presented in Figure 7, our GHPV model consists of three layers, the top layer (center circle) is the solely movie node for all videos within this movie or sitcom serial; the middle layer consists of place nodes (triangles), each of which represented a place; the leaf layer contains view nodes (shot sets), each of which represents a near-duplicated background view, and is associated with one place in the middle layer. Each view node has an inversed indexing list that records shots that contains similar key frames to this view. Using our GHPV model, the shots are hierarchically reversed indexed for efficient retrieval.

Figure 7: Graph-based Hierarchical Place-View (GHPV).

3.3 Efficient Search Large-Scale Database

Based on the hierarchical inversed indexing structure, the search process could be efficiently conducted with independency to the database volume. In traditional search process without inversed indexing, the computational cost is usually proportional to the data volume at least (For instance, larger that $o(n)$ is we have n shots). Based on our proposed GHPV model, we present an **A**pproximate **N**earest **N**eighbor search (**ANN**) strategy. Using ANN, the computational cost is largely reduced as follows:

For each key frame in user-uploaded query shot, we traverse the middle-level place layer in GHPV model, in this phrase, the Greedy N Best Paths (GNP) [5] algorithm is adopted for the purpose of leveraging search precision and efficiency. After this step, the top n best match places are further extended into the view layer retrieval to search within the view nodes, in which only the candidate shot that are indexed with the corresponding nodes are considered. Finally, in this subset, we pick the top m most similar place shots about this key frame.

The results of each top-rank shot lists are fused together to generate the final decision using Equation 6:

$$Rank_j = \sum_{k=1}^{K} \sum_{k'=1}^{K'_j} \| KF^k - KF_j^{k'} \|_{L2} \tag{6}$$

in which $Rank_j$ denotes the rank results of shot j^{th} shot in the GNP candidate shot set; K is the number of total key frames within the query shot; K'_j is the number of key frames for j^{th} shot in the GNP candidate shot set; KF^k is the pyramid-based spatial-encoding correlogram for k^{th} key frame in query shot, $KF_j^{k'}$ denotes the pyramid-based spatial-encoding correlogram for k'^{th} key frame in j^{th} shot; $\| \|_{L2}$ represents the L2 distance metric.

4. EXPERIMENTAL RESULTS

4.1 System Framework

We build a demo system *VISR (**VI**lab **S**cene **R**etrieval system)* for result validation, which enables fast online scene query from video database. Figure 8 shows the system framework. *VISR* has two basic functions, one is incremental indexing and the other is place query. The system video set is firstly built based on a original video set, by adding new videos or query shots, the system video set can index more and more scene and places.

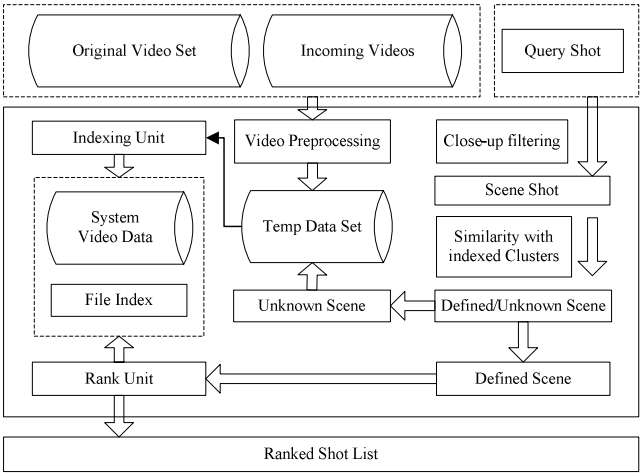

Figure 8. System framework of *VISR*.

A typical query is performed as follows: User uploads a video shot containing target place to our server. Key frames are online extracted from this query shot and the corresponding background masks are obtained. After feature extraction, the system retrieves shots with similar place and then returns a final ranked list. The results are presented in downright sub window of the *VISR* UI. The system framework of *VISR* demo is described as Figure 9.

Figure 9. Snapshot of *VISR* user interface.

Incremental Indexing Character: Our proposed framework consists of both offline part and online part. In offline part, the system will automatically examine the size of Data Buffer, when the size is bigger enough, the system will put these shots in Data Buffer into indexing unit for graph-based place view analysis and create new place in system index. By adding new videos or performing more queries, the system could refine the hierarchical place-view model incrementally. In online part, users can upload query shots, the system will judge whether the shots belong to a defined place in the system index. If the shot belongs to one defined place, it will be used by rank unit for shots ranking. If the distance between query shot and each existed cluster center is too large (lager than a given threshold), then this shot will be added to Data Buffer for further processing.

4.2 Experimental Database and Evaluation

In our experiments, we use three seasons (5~7) of *Friends* soap opera as our evaluation dataset. All videos are segmented into single camera shots, and then close-up filtering is performed to automatically select scene shots. Key frames of each shot are

extracted by simply correlogram comparison method. Over 500 shots are selected as the original data set for initial system place-view model generation. After incremental indexing, there are totally over 2,000 shots on our video database, which are then segmented onto over 5,000 views. Each key frame has a resolution of 300×480 pixels.

For performance evaluation, we select four representative places to form the query set, including Monica's apartment, Joey's apartment, "Central Perk" coffee house and a gallery between Monica and Joey's apartment. All of them are frequently appeared in *Friends* sitcom during these three seasons. Table 2 shows some information of the four typical queries.

Table 2: Information of the query places

	View point	Content	People	Room
Monica's Apartment	many	normal	normal	large
Joey's Apartment	normal	normal	normal	normal
Central perk Coffee house	normal	complex	many	large
Gallery	few	simple	few	small

For each type of place, we pick over 20 query shots, the ensemble of which forms a total 100-shot query to test the performance of our place retrieval algorithm. Among them, different actors are appeared within identical places and usually there are over one actors in shot frames with both object motion and global motion, which increase the challenge of our task. Figure 10 shows some of the queries.

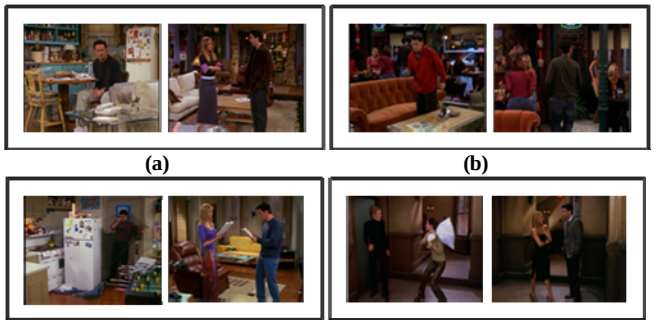

Figure 10: Four types of Queries. (a) Monica's apartment (b) Centrol Perk Coffee house (c) Joey's apartment (d) Gallery

We use Mean Average Precision at N (MAP@N) to evaluate system performance. This evaluation measure is commonly used in evaluating the ranking performance of information retrieval systems. MAP@N represents the mean precision of batch queries, each of which reveals its position-sensitive ranking precision within N^{th} position. Given n queries, MAP@N is defined as Equation 7:

$$MAP@N = \frac{1}{q}\sum_{q=1}^{Q}\sum_{r_q=1}^{n_q}\frac{P(r_q)\theta(N - pos(r_q))}{N} \qquad (7)$$

in which n is the number of query, $r_{\square}$ is a rank of $\square^{\square\square}$ query at position from 1 to N, $pos(r_{\square})$ is its position, $\theta()$ is Heaviside function: $\theta(\square) = 1$, if $x \geq 0$, and $\theta(\square) = 0$, otherwise. $P(r_q)$ is its precision; N is the number constraint of returned shots. The positive answer is pre-labeled for these queries.

4.3 Results and Discussions

Three experiments are conducted to evaluate the performance of our proposed method: 1) Histogram based similar shot retrieval. 2) Graph based place-view model. 3) Graph based place-view model with attention driven foreground removal. Figure 11-15 present the place retrieval results of the above methods. According to Table 2, there are four main factors we concerned, View Point, Content, People and Room, which will challenge the performance of place retrieval system. The four typical queries are the most representative cases.

Figure 11 shows the results of query (a) Monica's apartment, in which the MAP@10 for each method are 67%, 73% and 91%. Figure 12 is the result of query (b) Joey's apartment, in which the MAP@10 are, 67%, 70% and 83%. In Figure 13, the MAP@10 results of query (c) "Central Perk" coffee house are 91%, 96% and 98%. Figure 14 shows the results of query (d) Gallery, in which the MAP@10 is 51%, 62% and 73%.For each typical query, our proposed method outperforms over 5 to 18 percents in mean average precision.

Monica's apartment has many view points, its visual contents are not complex, and usually 1-5 peoples appear in the camera. Multi-view problem can be solved by utilizing Graph-based Place-view Model and attention-driven foreground removal can effectively deal with the main actors. Our proposed method is extremely adapted to this case, and the performance is actually much better than traditional methods. Compared with query (a), query (b) Joey's apartment is much smaller, and there are much less view points. In such cased, the three methods have similar precision. The average precision went down because there is one view in Monica's apartment which is quite like the main view of Joey's apartment. Query (c) "Central Perk" coffee house is a public place, which means lots of people and complex content. However, all methods have good performance because the appearance of this kind of place is much more stable than other places in movies. Query (d) Gallery is the simplest place in all queries. Fixed view point, bald background, small room and few people. However, due to all above mentioned features, the result of query (d) is greatly influenced by foreground objects. For this case, we utilize Attention-driven foreground removal to reduce the negative influences of the foreground objects. Although our foreground removal algorithm is not very precise, it's efficient enough in this specific application.

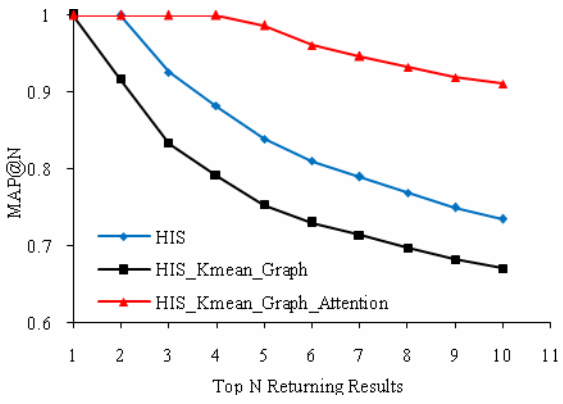

Figure 11. MAP of Query (a) Monica's apartment

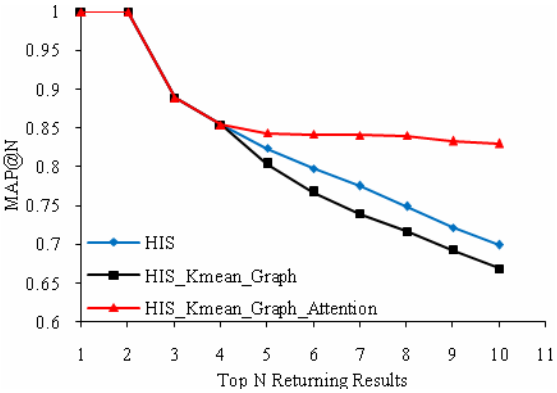

Figure 12. MAP of Query (b) Joey's apartment

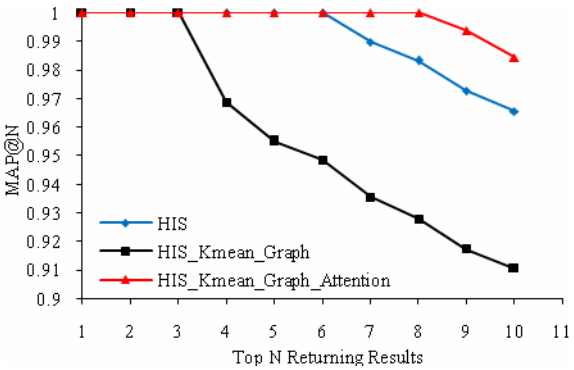

Figure 13. MAP of query (c) "Central Perk" Coffee house

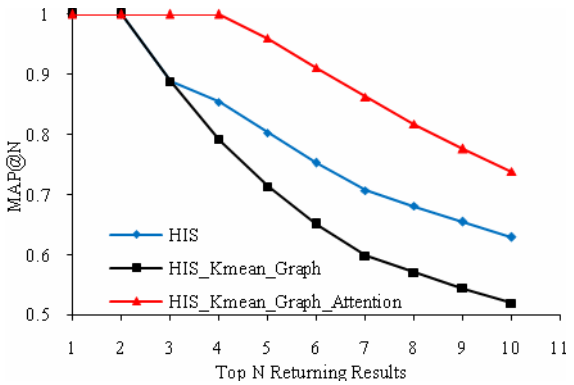

Figure 14. MAP of query (d) Gallery

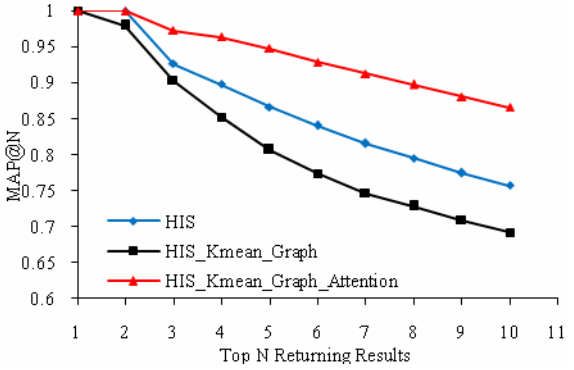

Figure 15. MAP of the complete query set

MAP results of the complete query set are given in Figure 15. MAP@10 for each method are 69%, 72% and 87%. Our proposed method has a higher average precision compared to the other methods. By using hierarchical inversed index, our system runs 2.3 times faster than traditional system. These experiments demonstrate the efficiency and robustness of our system.

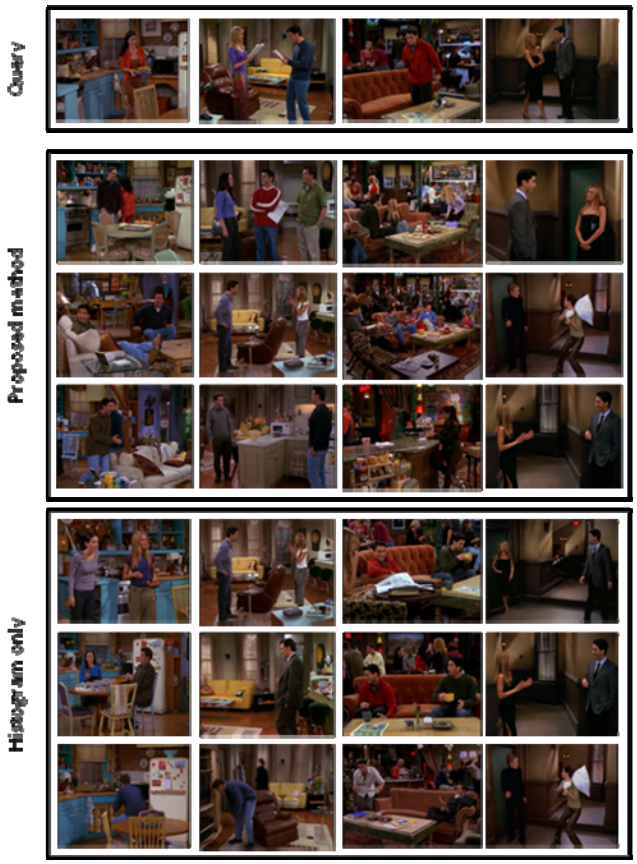

Figure 16. Some of the query results, each volume means a query with its results both using our proposed method and histogram matching. First volume: Monica's apartment; Second Volume: Joey's apartment; Third Volume "Central Perk"coffee house; Forth Volume: Gallery. Noticing that the returning results of our proposed method are with more semanticaly meanings, while still within the same place. And the histogram matching results are very similar in views, but lack of flexibity and in many cases would result in lower recall in searching "place".

Some of the query results are shown in Figure 16. We find that the proposed method can retrieve places in more views than traditional methods. Graph-based place-view model insurances the retrieval speed and precision. Attention-driven foreground removal enables selective comparison of the backgrounds. The combination of these two modules gives the entire system not only efficient search but also robust rank.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a novel place retrieval framework for efficient retrieval of physical-connective place views from movie and sitcom videos. Comparing with state-of-the-art, our method merits in selectivity, robustness, precision, and efficiency. Selectivity lies in the fact that our method adopts face-detection in close-up shot removal and leverages spatial &

motion saliency map in foreground removal; robustness lies in that we present a pyramid-based spatial encoding color correlogram (PSEC) in place appearance modeling; precision results from our Graph-based Hierarchical Place-View (GHPV) model that utilizes graph-based view connectivity analysis and normalized cut based place generation; efficiency mainly raises from the hierarchical inversed indexing to largely reduce computation complexity in search process. Our experimental validation over 36-hour *Friends* soap opera video database demonstrates the advantage of our proposed framework with state-of-the-art algorithms.

6. ACKNOWLEDGMENTS

This research is supported by State 863 High Technology R&D Project of China (No. 2006AA01Z197); the Program for China New Century Excellent Talents in University (NCET-05-03 34), Natural Science Foundation of China (No. 60775024).

7. REFERENCES

[1] YouTube: www.youtube.com

[2] Blinkx: http://www.blinkx.com

[3] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, Ramesh Jain, " Content-Based Image Retrieval at the End of the Early Years", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.12, December 2000, pp.1349-1380.

[4] Li Fei-Fei and Pietro Perona. A Bayesian hierarchical model for learning matural scene categories. Vol. 2, pp. 524- 531, Computer Vision and Pattern Recognition, 2005.

[5] G. Schindler, M. Brown and R. Szeliski. City-scale location recognition, Computer Vision and Pattern Recognition, pp.1-7, 2007.

[6] C Siagian, L Itti, Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention, IEEE Transactions on pattern Analysis and Machine Intelligence, Vol.29, No.2, pp. 300-312, 2007

[7] D Gokalp, S Aksoy, Scene Classification Using Bag-of-Regions Representations, Computer Vision and Pattern Recognition, pp.1-8, 2007

[8] Zhao, YJ, Wang, T., Wang, P. et. al., Scene Segmentation and Categorization Using NCuts, Proceedings of the 2nd International Workshop on Semantic Learning Applications in Multimedia, in association with Computer Vision and Pattern Recognition, Minneapolis, MN, 2007.
of Computer Vision and Pattern Recognition, pp.1-7.

[9] P Viola, MJ Jones, Robust Real-Time Face Detection, International Journal of Computer Vision, 2004.

[10] W.-H. Cheng, W.-T. Chu and J.-L. Wu, "A visual attention based region-of-interest detection," IEICE Transactions on Information and Systems, Vol.E88-D, NO.7, PP.1578-1586, 2005.

[11] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.20, No.11, pp. 1254-1259, Novermber, 1998.

[12] J Yuan, J Li, F Lin, B Zhang, "A unified shot boundary detection framework based on graph partition model", ACM SIG Multimedia, pp.539 – 542, 2005.

[13] C. Stauffer and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, vol. 2, 1999, pp. 246-252.

[14] Yueting Zhuang, Yong Rui, Huang, T.S., Mehrotra, S., Adaptive key frame extraction using unsupervised clustering, International Conference on Image Processing, pp.10-8186-8821-1/98, 1998.

[15] Jianbo Shi, Jitendra Malik, Normalized Cuts and Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.8, pp. 888-905, 2000.

[16] Huang J, Image indexing using color correlograms, IEEE Conference on Computer Vision and Pattern Recognition, 1997,762-768