THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Multimedia Descriptive Long Take Video Pool Retrived from One Shot Video Composition

Priya A.
B.Tech, Department of IT, SKP Engineering College, India
Ranjitha E.
B.Tech, Department of IT, SKP Engineering College, India
K. Ramadevi
Assistant Professor, Department of IT, SKP Engineering College, India

Abstract:
A large amount of short, single-shot videos are created by personal camcorder every day, such as the small video clips in family albums, and thus a solution for presenting and managing these video clips is highly desired. From the perspective of professionalism and artistry, long-take/shot video, also termed one shot video, is able to present events, persons or scenic spots in an informative manner. This paper presents a novel video composition system "Video Puzzle" which generates aesthetically enhanced Long- shot videos from short video clips. Our task here is to automatically composite several related single shots into a virtual long-take video with spatial and temporal consistency.

Key words*: Image retrieval, one-shot video, video authoring and video transition*

1. Introduction

With the popularity of personal digital devices, the Amount of home video data is growing explosively. These digital videos have several characteristics: (1) compared with former videos recorded by non-digital camcorder, nowadays videos are usually captured more casually due to the less constraint of storage, and thus the number of clips is often quite large; (2) many videos may only contain a single shot and are very short; and (3) their contents are diverse yet related with few major subjects or events. Users often need to maintain their own video clip collections captured at different locations and time. These unedited and unorganized videos bring difficulties to their management and manipulation. For example, when users want to share their story with others over video sharing websites and social networks, such as YouTube.com and Facebook.com, they will need to put more efforts in finding, organizing and uploading the small video clips. This could be an extremely difficult "Puzzle" for users. Previous efforts towards efficient browsing such large amount of videos mainly focus on video summarization. These methods aim to capture the main idea of the video collection in a broad way, which, however, are not sufficiently applicable for video browsing and presentation.

In this paper, we further investigate how to compose a content-consistent video from a video collection with an aesthetically attractive one-shot presentation. One-shot videos or long-shot video, also known as long-take video (we will exchange use them hereafter), means a single shot that is with relatively long duration. Long shot has been widely used in the professional film industry, MTV video2 and many other specific video domains owing to its uniqueness in presenting comprehensive content in a continuous and consistent way. However, capturing a high-quality long-shot video needs an accurate coordination between the camera movement and the captured object for a long period, which is usually difficult even for professionals. In this paper, we introduce a scheme, "Video Puzzle", which can automatically generate a virtual one-shot presentation from multiple video clips. Given a messy collection of video clips, Video Puzzle can select a clip subset with consistent major topic (similar with finding the clues and solving the Puzzle Games among the images [16]). The topic can refer to a person ,object, or a scene here. It can be specified by users or found with an automatic discovery method. The start-end frame correspondences of these clips are then established with an efficient coarse-to-fine method, and we compose them into a long clip in a seamless manner accordingly, i.e., a one-shot presentation. Therefore, Video Puzzle provides a novel presentation of video content that enables users to have a deeper impression story within the video collection. Fig. 1 shows the working process of Video Puzzle via two examples. The system can automatically discover video clips with "similar/continuous topics" in a video album and naturally stitch them into a single virtual long-take video, which can yield a cohesive presentation and convey a consistent underlying story. It is challenging as 1) it is generally hard to find shots which can be naturally combined among a large amount of candidate videos, and 2)generating seamless transition between video shots is difficult usually.

Figure 1

2. Related Work

It is the first work that proposes to compose coherent presentation automatically if there are appropriate domain-specific metadata associated with video segments and the composition techniques are established. Another preliminary work is [17]. The system automatically selects home video segments and aligns them with music to create an edited video segment which is quite different from ours. Our system concentrates on how to provide consecutive smooth video while theirs try to fit the video segment with the music. Other approaches [37] also end classify the segments by film theory, and compose them into a story. However, the target of the method is for professional videos which capture the whole story of a certain event while home videos and web videos often have no fixed single story.

2.1. Video Summarization

Many previous works focus on producing effective video summarization with visual friendliness and in a compact form. Existing methods can be classified into two categories, i.e. dynamic representation and static representation. Dynamic representation generates a video sequence that is composed of a series of sub-clips extracted from one or multiple video sequences or generated from a collection of photos, whereas static representation generally generates one or multiple images from video key-frames to facilitate not only their viewing but also transmission and storage. Although video summarization can reduce the cost of video browsing, there is a risk of missing detail sand the possibly inaccurate summarization also may cause inconvenience in browsing. For static representation, We propose a video management system for generating story board and also propose a comic-like video summarization algorithm. A morphological grouping technique is for finding 30 regions of high activity or motion from a video embedded in an image plane. The representation of motion in static images is a complex task with roots in art and science . For dynamic representation, We provide a scenario-based dynamic video abstractions using graph matching and also propose a hierarchical technique to identify clinically relevant segments in diagnostic copy videos and their associated key-frames, and then create a rich video summary. This approach is adaptive to video contents, and it represents the clinically relevant video segments hierarchically to facilitate fast video browsing. These narratives are characterized for being compact, coherent and interactive. This system can be used to create interactive posters for video clips.

.

2.2. Video Editing/Composition

Our work is also related to video editing and composition. In comparison with still image editing, content-based video editing faces the additional challenges of maintaining the spatial-temporal consistency with respect to geometry. This brings up difficulties of seamlessly modifying video contents, such as inserting or removing an object. We provide a solution based on an unsupervised inference of view-dependent depth maps for all video frames. Transfer desired features from a source video to the target video such as colorizing videos, reducing video blurs, and video rhythm adjustment. Recently, we have studied automatic broadcast soccer video composition. There also exist studies on video texture which aims to provide a continuous and infinitely varying stream of image. We need to mention that there exist several media composition works that all extract a media subset through finding a path in a graph constructed by media samples, but as such a major contribution of our work is the coarse-to-fine process for identifying the corresponding video clips (i.e., the media graph construction process)The overall scheme “Video Puzzle” aims to discover content-consistent video shots and composes them into a virtual long-take video. To this end, we propose a novel graph-based visualization and path finding approach. The graph is constructed based on geometry matching (homograph mapping) and object matching (human, face). Based on the multi-cue content matching, the transition of video shots becomes meaningful and seamless.

3. An Overview of the Scheme

Our task is to automatically compose several related video shots into a virtual long-take video with spatial and temporal consistency, and it is different from the traditional works that try to either find a group of similar video clips or fit the composed video with extra information such as music or metadata. For a given video collection that contains video clips. The system mainly contains three key

components. Firstly, we implement a coarse-to-fine partial matching scheme to generate a matching graph of the video collection. The matching scheme serves as a three-level matching, i.e., video pair selection, sequence-sequence correspondence finding, and frame-level exact matching. The video pair selection acts as an evidence for ensuring the *non-redundant* and *complete* quality of the generated one-shot video. It uses a hashing-based method to quickly

obtain the video similarity measurement. We then find sequence correspondence of the selected video pairs through local key points matching. The final frame-level matching aims to find different matched objects to provide variant and rich clues for video transition generation. We implement three object matching methods in this part, i.e., salient object matching using local visual pattern discovery and human and face appearance matching based on automatic human and face localization. Secondly, we design a flexible scheme to select the optimal video compositions from a constructed video matching graph. The video selection task turns out to find the longest path in the graph by constructing a video matching graph following three criteria, i.e., continuity, completeness and diversity. This selection scheme can either work fully automatically by creating one-shot videos with globally optimal content consistency or work interactively with users by generating one-shot videos with optional topics (such as the specified key objects or persons).Finally, we compose the video correspondence pair one by one. We propose a space temporal morphing-based transition through matched local patterns, i.e., matched local common pattern, matched human face.

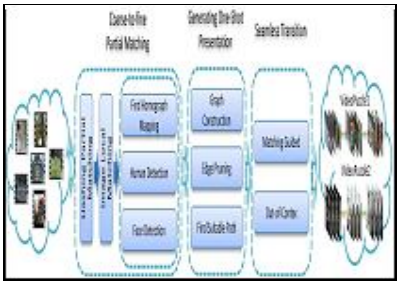

Figure 2

4. Coarse-to-Fine Partial Matching

In this section, we introduce the first component of the system, namely, coarse-to-fine partial matching. The target of this part is to (a) produce video similarity measurement acting as evidence for ensuring the *non-redundant* and *complete* quality of the generated video; (b) fast and accurately locate the sequences in video pairs with start-end content correspondence and (c) find the key frame pairs with transition clues in the correspondence sequences. We first use a hashing-based method to quickly obtain the video similarity measurement. Then we try to match two video sub-sequences in order to generate continuous transition. Finally, specific transition clues are obtained for video composition through local common patterndiscovery, human appearance modeling and face appearance modeling.

4.1. Hashing-Based Video Pair Selection

In this part, we adopt the recently proposed Partition Min-Hashing algorithm to rapidly calculate the frame

Partial similarity between every pair of videos and the computed frame similarity is accumulated to estimate the video similarity measurement. Then, the video pairs with high similarity are selected as candidate pairs to generate one-shot videos. A graph of video similarity is built based on the results of video pair selection. In practice, we filter out most video pairs with low similarity and only retain up to four video pairs as the matching candidates for each video. Therefore, the computational cost for the further video matching steps is largely reduced.

4.1.1. Min -Hashing

In the min-hash algorithm, a hash function is applied to all visual words in an image without considering their locations, and the visual word with minimum hash value is selected as a global descriptor of the given image. When an image is represented by a set of visual words, the similarity between two images can be defined as the similarity between the two corresponding sets of visual words i.e.,, which is simply the ratio of the intersection to the union of the two set. .Min-hash is a hash function ,which maps a set to a value . More specifically, a hash function is applied to each visual word in the set , and the visual word that has minimum hashed value is returned as the min-hash. The computation of the min-hash of a set involves the hash of every element in the set and the time cost thus scales linearly with the size of the set . In our case, we are interested in finding images which have similarity greater than a threshold.

.

4.1.2. Partition Min-Hash

However, unlike text documents which are usually represented with bags of words, images are strongly characterized by their 2D structured objects which are often spatially localized in the image. Partition min-hash is proposed as a novel hashing scheme to exploit the locality. In this image is first divided into partitions. Hashing is then applied independently to the visual words within each partition to compute a min-hash value. With evenly divided partitions, the duplicate may be split into two or more partitions.

4.2. Sequence Matching

In this subsection, we aim to accurately match two video subsequences within the selected video pairs in order to generate continuous transition. We propose a method that uses image local matching to get the correspondence of two sub-sequences.

4.2.1. Image Local Key points Matching

We use SIFT Color Moments with Difference of Gaussians (DOG) key point detector. Existing studies demonstrate that the SIFT descriptors and Color Moments are complementary to each other, one describing the local structure and the other providing higher

order information of local differences. We concatenate these two features to describe each local key point. Given two frames (the source image, frame in video , and the target image, frame in video), the best candidate match for each key point of the source image is found by identifying its nearest neighbor among the key points from the target image. Then nearest neighbor is defined as the key point with the minimum Euclidean distance. Since there will be many key points from the source image that do not have any correct match in the target image, such as those that arise from background clutter are not detected in the target image, it is useful to discard them.

4.2.2. Frame Similarity to Sequence Correspondence

To locate the sequence correspondence of two videos, we sample the video frames in a constant rate. Given two frame sequences and of same length in two videos, we first calculate the maximum similarity over the frames. For each video pair and , we obtain the sequence correspondence by finding the sequence pair with the largest sequence similarity. The video similarity in the graph is also replaced by this sequence similarity. This process of searching for sequence correspondence is critical in our system. It determines whether video clips can be composed with the other videos

.

4.3. Transition Clues for Video Composition

Given two matched sequences, we select the frame pairs as the transition key frames according to several transition clues.

4.3.1. Cross-Frame Common Pattern Discovery

The first transition clue we use is based on image matching. Since image matching often contains a large amount of outliers, we need a robust fitting method to find the common pattern. Specifically, common pattern denotes those matching pairs that share the same or similar homogeneous transformation parameters.

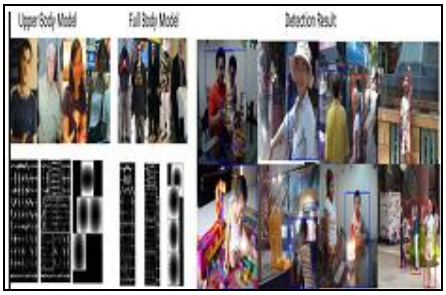

Figure 3

4.3.2. Human Appearance Matching

The frames from two videos are also matched according to the appearance of human contained in the video. Firstly, automatic human body detection is accomplished. The part-based detection model contains two parts, one describing full view (denoted as root model) and the other describing part views (denoted as part models). The appearances of the detected human bodies are represented in color histogram and matched to find the same person appeared in different frames videos.

4.3.3. Face Appearance Matching

We also implement the-art multi-view face detector and active shape model for face alignment. For each frame, we perform the near-frontal face detector to localize the face area as well as several facial parts, such as eyes, mouth, nose and face. A frame with face is assumed to be matched with another frame with face according to the following criteria:

- Both face areas should be large enough. Small face areas are much less important since video matching and transition on small area frequently lead to unnatural effects. In our implementation, we set the threshold to 3,600 pixels.
- The faces should belong to the same person. We first perform the face alignment procedure to align the faces and then calculate the Euclidean distance for the feature vectors extracted from each face pair. A threshold is empirically set to remove most mismatched candidates.
- The two face poses should not very much. The output of the face detector [18] includes the pose view information.

Figure 4

Based on the above three transition clues, we locate candidate key frame pairs from two sequences as follows:

- Frame pairs with the same object. For each pair of frames in the sequence correspondence, we perform common pattern discovery, and the frame pair with the maximal pattern support is then chosen as a key frame pair.

- Frame pairs with the same person. If the matched score of persons/faces within a frame pair is greater than a predefined threshold, the frame pair is chosen as a key frame pair.

5. Generation of One-Shot Presentation

Given a video collection that contains video clips. We construct a matching graph where denotes the video, is the directed weight for then and , which is the maximal sequence matching score of the video and Our task here is to find a path from the directed graph to connect the short shots into a long-take Shot.

- **Continuity**: Each edge on the path should have a weight greater than a predefined threshold. Otherwise, the edge is removed.
- **Completeness**: The overall path should be sufficiently long. To ensure the completeness of the video, large number of combined clips is preferred.
- **Diversity**: The nodes should have large variety. Since the matched clips possibly contain many near-duplicate versions, we need to exclude them to retain the compactness of the composited video. This step is accomplished by exploring the similarities among videos.

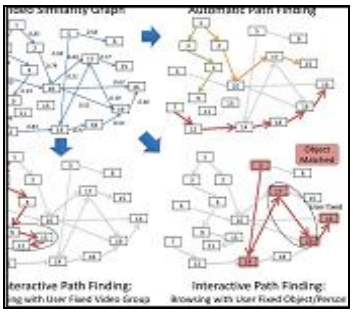

Figure 5

5.1. Edge Pruning

We prune the cycle paths in the graph to avoid the repeated clips in the composed video. An important and also the most straight forward criterion is time constraint. People are used to watching videos in the order of time, especially for home video browsing. We use the timestamp metadata of the video clips to ensure that the shots maintain the temporal relationships in the composition process. However, we also notice that many video clips lack of such metadata. Therefore, for those video clips, we need to design extra content-based edge pruning method to reduce the cycle graph. Here we use Depth-First-Search to detect all the nodes that have a cycle in the graph. We can locate the edges within the cycle, then the edge with lowest weight will be pruned.

5.2. Path Finding

5.2.1. Automatic Path Finding

The maximal paths can be found automatically. After finding the longest path over the graph, all the edge weight linking to those nodes in the path should be scaled by a factor (in our implementation) to reduce the possibility for these nodes to be selected again. We then find the longest path again in the updated graph. This procedure can be iterated until reaching the criterion that the sum of weights in the final path is less than a threshold.

5.2.2. Interactive Path Finding

For personal usage, the technique can help to find and composite consecutive video clips with human interaction. A user may expect a one-shot video that contains a specified key video clip or focuses on a specific object or scene.

- **User fixing one video clip** : We find the maximal path from other nodes to node and the longest path from node to other nodes. The overall one-shot video is then generated with the combined path. Since the constructed graph is an acyclic graph, it is guaranteed that no node will be selected more than once.
- **User fixing a group of video clips:** User may want to fix several similar video clips into the composition. First, the group of video clips is deemed as a virtual node. All the edges linking to the group clips is linked to this node. The problem then turns out to find the path passing through the node.
- **User fixing the matched object:** We can list many matched objects within the video album, such that user can select a matched object. To find a one-shot video that contains this matched object, we first locate the two video and that contain the selected object, and the problem then turns to finding the longest path that ends at the node and starts at the node.

6. Seamless Video Composition

Here we introduce how to compose the selected video clips into one-shot video and the key problem is to smooth the visual discontinuities at the transitions. For each two best matched frames, all the matches are local, such as common patterns, human bodies, and faces. Since directly stitching the two videos based on these two frames may lead to abrupt change, we need to consider adding natural transition, which act as the link between the two consecutive videos, in the final virtual long-take video.
we use the following procedures to generate the more natural transition between videos:

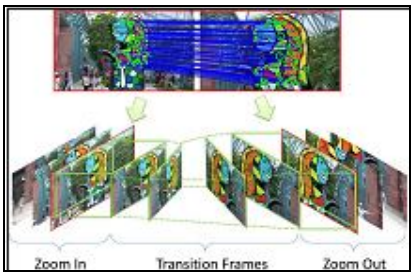

Figure 6

- **Finding the minimum matched area:** The transition between two matched objects often needs to be smooth and continuous. Thus, instead of simply generating a transition between the frames and in the video and , we generate a transition between the largest matched sub-windows and . We select the matched areas by taking three factors in consideration to guarantee the smoothness: (1) keeping the width height ratio of sub window; (2) finding the minimum matched area covering most of local matched points; and (3) the offset between the centers of the local matched points within and should be minimized.
- **Focusing on the matched object:** After locating the and, We find frames before the frame in the video and produce Zoom-In effect in the sequence. Similarly, Zoom-Out effect is produced on video .
- **Feathering on unmatched area:** The transition and may still have effect for the unmatched area. To address this issue, we adopt the feathering approach commonly used for image mosaics. That is, we weight the pixels in each frame proportionally to edge and their distance to the matching points center.

7. Experiments

7.1. Dataset Preparation

Several experiments were performed to verify the effectiveness of the proposed "Video Puzzle" framework. Three video albums are prepared for the experiments. We denote them as , and , where and are typical home video albums and the videos in are collected fromYoutube.com with some keywords of famous landmarks. The set contains 68 video clips. They are captured in a trip and the locations vary widely, including beach, landscape, and woods road. The set contains 186 video clips that record the birthday parties of a child, piano practice scenes.

7.2. Implementation Details

For each video album, we first construct a bag-of-words(BOW) model using SIFT features for the partition minimum hashing The number of features per image ranges from 200 to 1000,and we have quantized them using a visual word vocabulary with one million visual words. It takes about 100 ms per image to extract the features.


Figure 7

2. Each image is divided into about 100 partitions with 50%overlap as recommended in [24]. We uniformly sample 1/5 of the frames to accelerate the video similarity measurement. The frame similarities are accumulated to form video similarity. We then set a threshold to make the similarity graph sparse. For each connected video pairs, local matching-based sequence matching is performed. Finally, exact matching frame pairs are located.

7.3. *One-Shot Video Generation*

In this section, we give some generated one-shot examples from the video album .

7.3.1. Transition Effect Evaluation

We first check the effect of the generated transitions. We compare the proposed transition method with the widely-used method, i.e., fade-in and fade out. Given two corresponding frames, we compare the generated results.

.

7.3.2. Family Video Browsing

By taking the as an example, we show the video matching graph after edge pruning in Each node represents a video in . The edge linking two nodes indicates a match between them. The matched region or object is shown upon the edge with a confidence identifying the sequence similarity. The matching graph brings a new tool for video album browsing. Unlike the traditional video summarization methods which try to extract key content of video album and present in an concise manner, the matching graph has several characteristics:

- It presents the video album in a wide-range manner and gives the user the direct impression over the large numbers of clips.
- More importantly, the matching graph gives the correspondence information of two videos. The linking edge introduces the browser to explore the whole set with continuous content rather than one by one.
- It also highlights the "key" clip in the video album .The "key" clip has the most number of edges connecting from and to it.

7.3.3. Landmark Videos From Websites

We also evaluate our method on , which contains many online video clips about "Eiffel Tower". These two scenes are mutually acting as noise. As shown in Fig. 10, we output the longest path which is a presentation about "Eiffel Tower", composed with 5 shots. The match clue used here is induced by common pattern discovery. Usually, for each frame, about 1000SIFT features can be extracted. However, the images contain a lot of outliers. The traditional RANSAC method fails in this case while our method still can find the common pattern.

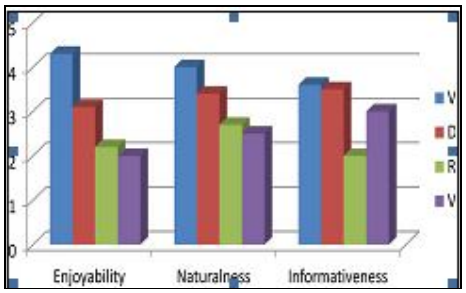

Figure 8

7.4. User Study

To subjectively evaluate the results, we compare the following methods: (1) Our "Video Puzzle" with content-based transition (denoted as "VP"); (2) The video produced by direct compositing video clips discovered by the "Video Puzzle" scheme (denoted as "DC"); (3) The automatically edited videos with the videos produced by connecting randomly selected video clips (denoted as "RS"). (4) The video summarization method proposed in [28] (denoted as "VS") groups all video shots into scenes and then generates a skimming for each scene by a graph-based mining method. Then, the skimming are concatenated into a skimming of the whole video set. The three sets of videos are used to produce the baselines and the "Video Puzzle". Totally, we extract 8 video puzzles that are composed by 35 video clips. Accordingly, we also composite8 DC videos with the same clips. Another 8 RS videos are randomly generated within each dataset. 20 evaluators were invite do participate in the user study.

.

7.5. On the Robustness of Our Approach

In order to test the robustness of our approach, we performance other test on the three albums. For each album, we randomly selected 20 video clips from the other two albums and added them to the target album as noisy data. We then generated the one-shot videos again. Interestingly, we find that the generated one-shot videos is exactly the same with our previous results. Therefore, the robustness of the system is acceptable although there exist several mismatching cases.

7.6. Limitation and Future Work

Overall, the Video Puzzle scheme performs well on these three albums. It can automatically discover the consistent topics within personal albums or online landmark albums. It generates consistent video composition based on the semantic matching. However, it also has several limitations:

- The proposed method only works when the number of videos is large enough and there contains certain consistent topics.
- The transition clues used in this scheme, i.e., the common key points pattern and face/human matching, may produce false alarm.

.

8. Conclusion

In this paper, we proposed "Video Puzzle", an integrated system for both video summarization, browsing and presentation, based on large amount of personal and web video clips. This system automatically collects content-consistent video clips and generates an one-shot presentation using them. It can facilitate family album management and web video categorization. We demonstrated two example applications using "Video Puzzle" and the results show that it has great potential to beused in future video management systems.

9. References

1. C. Wang , C. Xu , E. Chng and H. Lu "Automatic composition of broadcast sports video", Multimedia Syst., vol. 14, no. 4, pp.179 -193 2008
2. C. Huang , H. Ai , Y. Li and S. Lao "High-performance rotation invariant multi view face detection", IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 4, pp.671 -686 2007
3. C. C. Nikolaidis "Video shot detection and condensed representation. A review", IEEE Signal Process. Mag., vol. 23, no. 2, pp.28 -37 2006
4. T. Do , K. Hua and M. Tantaoui "P2VoD: Providing fault tolerant video-on-demand streaming in peer-to-peer environment", Proc. 2004 IEEE Int. Conf. Communications, vol. 3, pp.1467 -1472 2004
5. J. Liu and M. Zhou "Tree-assisted gossiping for overlay video distribution", Multimedia Tools Appl., vol. 29, no. 3, pp.211 -232 2006
6. S. Jin and A. Bestavros "Gismo: A generator of internet streaming media objects and workloads", SIGMETRICS Perform. Eval. Rev., vol. 29, no. 3, pp.2 -10 2001
7. J. Cao et al., &ldquo,A Multi-Agent Negotiation Based Service Composition Method for On-Demand Service,&rdquo, Proc. Int',l Conf. Services Computing (SCC), pp. 329-332, 2005.
8. J. Cao et al., &ldquo,A Multi-Agent Negotiation Based Service Composition Method for On-Demand Service,&rdquo, Proc. Int',l Conf. Services Computing (SCC), pp. 329-332, 2005.
9. A.Hampapur and R. M. Bolle Videogrep: Video copy detection using inverted file indices, 2001
10. J. Sivic and A.Zisserman "Video google: A text retrieval approach to object matching in videos", Proc. IEEE Int. Conf. Computer Vision (ICCV), pp.1470 2003