Video Puzzle: Descriptive One-Shot Video Composition

[1]Seetha.Mahesh, [2]P.Sai Prasad
[1]student, [2]Assistant Professor
[1]digital Electronics and Communication System
[1]siddartha Educational Academy Group Of Institutions, Tirupati, A.P,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. 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. In this project a novel video composition system "Video Puzzle" is proposed which generates aesthetically enhanced long-shot videos from short video clips. Automatic composition of several related single shots into a virtual long-take video is done. A novel framework is designed to compose descriptive long-take video with content-consistent shots retrieved from a video pool . For each video, frame-by-frame search is performed over the entire pool to find start-end content correspondences through a coarse-to-fine partial matching process. The content correspondence here is general and can refer to the matched regions or objects, such as human body and face. The content consistency of these correspondences enables us to design several shot transition schemes to seamlessly stitch one shot to another in a spatially and temporally consistent manner. The entire long-take video thus comprises several single shots with consistent contents and fluent transitions. Meanwhile, with the generated matching graph of videos, the proposed system can also provide an efficient video browsing mode.**

***Index Terms* - Image retrieval, one-shot video, video authoring, and video transition**

I. INTRODUCTION

With the popularity of personal digital devices, the amount of home video data is growing explosively. These digital videos have several characteristics: 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; Many videos may only contain a single shot and are very short; and 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 up loading 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 shot presentation. One-shot videos or long-shot video, also known as long-take video (we will exchange ably 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 video 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.

Here a scheme, "Video Puzzle" is introduced 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). 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 of the story within the video collection. The working process of Video Puzzle has 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

- It is generally hard to find shots which can be naturally combined among a large amount of candidate videos,
- Generating seamless transition between video shots is difficult usually.

The contribution of our work can be summarized as follows:

- The proposed scheme is able to extract video contents about a specific topic and compose them into a virtual one-shot presentation. The scheme is flexible and several components can be customized and applied to different applications.

- An efficient method to find the content correspondence so multiple videos and then compose them into a clip with an optimized approach is proposed. This scheme has two applications based on the video puzzle scheme, one about home video presentation and the other about landmark video generation specifically.

The two specific applications introduced are:

- Personal video presentation with a large set of personal video contents, a video matching graph is generated which explicitly shows the content-consecutive relation of videos. The storyline of the video album found by Video Puzzle will automatically pop up. Besides, user only needs to appoint a specific person or scene and then we can generate a one-shot presentation to describe the corresponding person or scene by mining the video grap.
- Comprehensive landmark video generation. With multiple web videos that describe the same landmark, we are able to generate a one-shot visual description of the landmark, which contains more comprehensive visual description of the landmark, such as the visual contents captured from different views.

II. PROBLEM DEFINITION

The existing system explored time flow manipulation in video such as the creation of new videos in which events that occurred at different times are displayed simultaneously and the content retrieval is only for the stationary objects. The organized long video provides user interested data with respect to time and location.

Limitations:

- Either Human or Object Categorization is achieved.
- It can only deal with static background for foreground extraction.
- A Risk of Missing Details.

III. OBJECTIVE

The main aim of this project is to create a long shot video from short video clips based on image content retrieval. There by the video retrieved can be presented in an informative manner. With the popularity of personal digital devices, the amount of home video data is growing explosively. These digital videos have several characteristics.

- Many videos may only contain single shot.
- Their contents are diversified.

The proposed system aims to automatically discover content consistent video shots. These videos compose in to a virtual long take video with spatial and temporal consistency. It also provides a tree structured collection for ease of video browsing. The contribution of our work can be summarized as follows: a scheme video puzzle is proposed. It is able to extract video contents about a specific topic and compose them into a virtual one-shot presentation.

The scheme is flexible and several components can be customized and applied to different applications. An efficient method to find the content correspondences of multiple videos and then compose them into a clip with an optimized approach is proposed. We introduce two applications based on the video puzzle scheme, one about home video presentation and the other about landmark video generation.

Personal video presentation with a large set of personal video contents, A video matching graph is generatedwhich explicitly shows the content-consecutive relation of videos. The storyline of the video album found by Video Puzzle will automatically pop up. Besides, user only needs to appoint a specific person or scene.

IV. EXISTING SYSTEM

Part Based Model for Object Detection

Object recognition is one of the fundamental challenges in computer vision. In this regard, there is a problem of detecting and localizing generic objects from categories such as people or cars in static images. This is a difficult problem because objects in such categories can vary greatly in appearance. Variations arise not only from changes in illumination and viewpoint, but also due to non-rigid deformations, and intra class variability in shape and other visual properties. For example, people wear different clothes and take a variety of poses while cars come in a various shapes and colors. An object detection system is described that represents highly variable objects using mixtures of multi scale deformable part models. These models are trained using a discriminative procedure that only requires bounding boxes for the objects in a set of images.

This approach builds on the pictorial structures framework Pictorial structures represent objects by a collection of parts arranged in a deformable configuration. Each part captures local appearance properties of an object while deformable configuration is characterized by spring-like connections between certain pairs of parts. Deformable part models such as pictorial structures provide an elegant framework for object detection.

It has been difficult to establish their value in practice. On difficult datasets deformable part models are often outperformed by simpler models such as rigid template. One of the goals of our work is to address this performance gap. While deformable models can capture significant variations in appearance, a single deformable model is often not expressive enough to represent a rich object category. Consider the problem of modeling the appearance of bicycles in photographs. People build bicycles of different types (e.g., mountain bikes, tandems, and 19th-century cycles with one big wheel and a small one) and view them in various poses.The system described here uses mixture models to deal with these more significant variations. We are ultimately interested in modeling objects using “visual grammars”.

Grammar based models generalize deformable part models by representing objects using variable hierarchical structures. Grammar based models allow for, and explicitly model, structural variations. These models also provide a natural framework for sharing information and computation between different object classes. For example, different models might share reusable parts. Although grammar based models are our ultimate goal, we have adopted a research methodology under which we gradually move toward richer models while maintaining a high level of performance. Improving performance by enriched models is surprisingly difficult. Simple models have historically outperformed sophisticated models in computer vision, speech recognition, machine translation and information retrieval. For example, speech recognition and machine translation systems based on n-gram language are used.

One reason why simple models can perform better in practice is that rich models often suffer from difficulties in training. For object detection, rigid templates and bag of features models can be easily trained using discriminative methods such as support vector machines (SVM). Richer models are more difficult to train, in particular because they often make use of latent information. Consider the problem of training a part-based model from images labeled only with bounding boxes around the objects of interest. Since the part locations are not labeled, they must be treated as latent (hidden) variables during training. More complete labeling might support better training, but it can also result in inferior training if the labeling used suboptimal parts. Automatic part labeling has the potential to achieve better performance by automatically finding effective parts. More elaborate labeling is also time consuming and expensive.

The first innovation involves enriching the Dalal- Triggs model using a star structured part-based model defined by a “root” filter (analogous to the Dalal-Triggs filter) plus a set of parts filters and associated deformation models. The score of one of our star models at a particular position and scale within an image is the score of the root filter at the given location plus the sum over parts of the maximum, over placements of that part, of the part filter score on its location minus a deformation cost measuring the deviation of the part from its ideal location relative to the root. Both root and part filter scores are defined by the dot product between a filter (a set of weights) and a sub window of a computed from the input image. Automatic part labeling has the potential to achieve better performance by automatically finding effective parts. More elaborate labeling is also time consuming and expensive. More complete labeling might support better training.

Facial Recognition Using Active Shape Models

It is an improved method for facial recognition of frontal faces using local patches around well defined facial landmarks. Our method aims at rectifying the problems of illumination variation and in-plane rotation of faces by only using specific discriminative areas on a face thus making it more robust. 79 landmarks are automatically fitted onto all faces in our training and test set using a pre-trained Active Shape Model. Local patches of fixed dimension are built around the most discriminative and accurate landmarks and then used to obtain features. It is these features that are used to differentiate one class from another using a Support Vector Machine as the classifier in a one against the rest form. We evaluate our scheme on random training and test sets drawn from two different databases can show that our method is capable of good recognition rates.

Facial recognition schemes are increasingly becoming more accurate; however, the combined effect of illumination changes, pose variations and in-plane rotations of subjects has been known to throw off the accuracy of several schemes. Our focus is on illumination and in-plane effects and we do not address the problem posed by pose variations in this paper. Several solutions have been proposed to deal with the problem of illumination. Such schemes include de-illumination and reillumination of faces in the image domain as described in illumination normalization using histogram equalization and using Near-Infrared images. All of the above schemes do achieve good results but focus mainly on compensating for illumination affects rather than using an approach that is inherently robust to it.

It has been shown that a local approach to face recognition is more robust to illumination effects than a global approach .It is for this reason that we focus on the use of features extracted from small two-dimensional (2D) regions around selected facial landmarks for our recognition algorithm. A modified Active Shape Model (ASM) is used to determine the locations of 79 landmark points across all faces in our training and test databases. Local patches are isolated around each landmark and used to build features unique to each class using a combination of Gabor filter banks and Principal Component Analysis (PCA). These features are used to train a Support Vector Machine (SVM) which serves as our classifier. Such a scheme harnesses a lot of information from every facial image unlike global approaches which utilize pixel intensities in the image as a whole and thus can suffer from noise in the background, in-plane rotations etc. Similar local approaches have been followed and use Active Shape Models to find landmarks of interest on a face and then compare the facial shape of a test image with those in a training database to classify the test image. Active Shape Models (ASMs) are aimed at automatically locating landmark points that define the shape of any statistically modeled object in an image. When modeling faces, the landmark points of interest consist of points that lie along the shape boundaries of facial features such as the eyes, lips, nose and mouth.

The training stage of an ASM involves the building of a statistical facial model from a training set containing images with manually annotated landmarks. The land marking scheme used by us consists of 79 facial points as shown in Figure 1. Our training set comprised of 500 images of 115 subjects from the query set of the still face challenge problem of the MBGC-2008 database. The shapes in the training set are aligned with each other using Generalized Procrustes Analysis(GPA) and then used to generate a mean shape of a typical face. Subsequently, statistical models of the grey level intensities of the region around each landmark are built using 2D profiles which are generated by sampling the image in a square region around each landmark. Such profiles are generated for each landmark point in each image and for four different levels in an image pyramid.

At the testing stage, the Open CV implementation of the Viola Jones face detector is used for locating the face in an image. Once the face has been detected, the mean face is scaled, rotated and translated using a similarity transform to roughly fit on top of the face in the test image. Multi-level profiles are constructed for the image in the same way as they were at the training stage. Landmarks are repeatedly moved into locations with profiles that best match the mean profile for that landmark until there is no significant change in their positions between two successive iterations.

This dissertation evaluates some variations of the Active Shape Model of Coots, as applied to monochrome front views of upright faces with neutral expressions. We evaluate our scheme on random training and test sets drawn from two different databases show that our method is capable of good recognition rates.

V. PROPOSED SYSTEM

Module Description

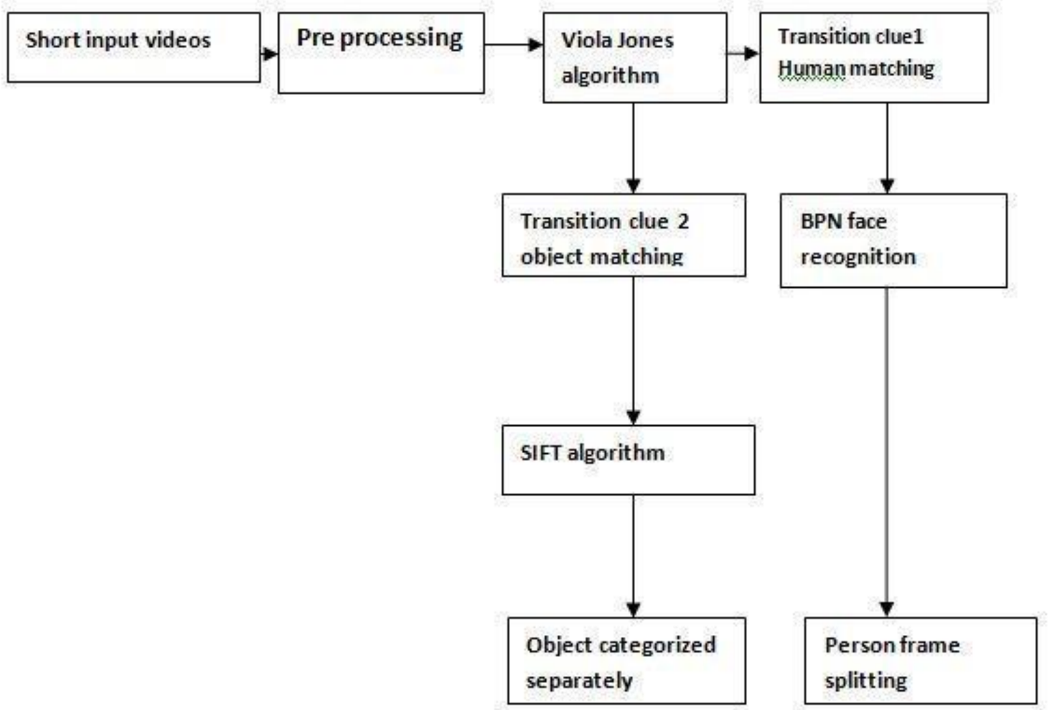

Fig 3.1 Block diagram of proposed system

- **Preprocessing -** First given Input short videos are converted into frames. Information less frames (Mean of Input frame<15) are eliminated. After that each frame is resized. Then all frames are merged into a single video for video categorization.
- **Categorization Based on Transition Clues -** Videos are categorized by using transition clues like human, object. Human clue is taken for first categorization by using Viola-Jones algorithm, if faces are not detected in frames those frames are separated into another process for object matching clue. After matching then it will apply separate algorithms for both object and person videos separately.
- **Video Composition -** Object & sequence matching process are done by using SIFT algorithm (Scale-invariant feature transform). Related Object frames and related sequence frames are categorized into separate folder respectively. Finally categorized frames are converted into separate videos.

Block Diagram Description

- **Converting Input Videos In to Frames**

First input videos with different resolutions are taken, after that all the videos are converted into frames by using Matlab syntax and total video converter and resize all the frames to equal resolution and finally all the frames are merged into single video. This merged video having equal resolution.

- **Viola Jones Algorithm**

The basic principle of the Viola-Jones algorithm is to scan a sub-window capable of detecting faces across a given input image. The standard image processing approach would be to rescale the input image to different sizes and then run the fixed size detector through these images. This approach turns out to be rather time consuming due to the calculation of the different size images. Contrary to the standard approach Viola-Jones rescale the detector instead of the input image and run the detector many times through the image – each time with a different size. At first one might suspect both approaches to be equally time consuming, but Viola Jones have devised a scale invariant detector that requires the same number of calculations whatever the size

The basic principle of the Viola-Jones face detection algorithm is to scan the detector many times through the same image – each time with a new size. Even if an image should contain one or more faces it is obvious that an excessive large amount of the evaluated sub-windows would still be negatives (non-faces). This realization leads to a different formulation of the problem: Instead of finding faces, the algorithm should discard non-faces. The thought behind this statement is that it is faster to discard a non-face than to find a face. With this in mind a detector consisting of only one (strong) classifier suddenly seems inefficient since the evaluation time is constant no matter the input. Hence the need for a cascaded classifier arises. The cascaded classifier is composed of stages each containing a strong classifier. The job of each stage is to determine whether a given sub-window is definitely not a face or maybe a face. When a sub-window is classified to be a non-face by a given stage it is immediately discarded. Conversely a sub-window classified as a maybe-face is passed on to the next stage in the cascade. It follows that the more stages a given sub-window passes, the higher the chance the sub-window actually contains a face.

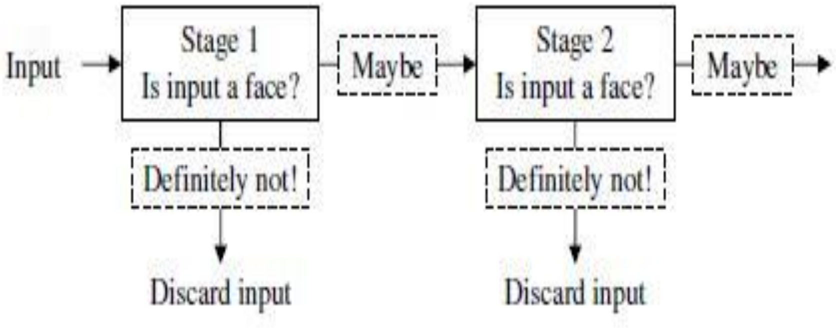

Fig 3.2 viola Jones cascaded classifier

In a single stage classifier one would normally accept false negatives in order to reduce the false positive rate. However, for the first stages in the staged classifier false positives are not considered to be a problem since the succeeding stages are expected to sort them out. Therefore Viola-Jones prescribes the acceptance of many false positives in the initial stages. Consequently the amount of false negatives in the final staged classifier is expected to be very small. Viola-Jones also refer to the cascaded classifier as an attentional cascade. This name implies that more attention (computing power) is directed towards the regions of the image suspected to contain faces. It follows that when training a given stage, say n, the negative examples should of course be false negatives generated by stage.

- **SIFT Algorithm**

The SIFT algorithm (Scale Invariant Feature Transform) proposed by Lowe [1]is an approach for extracting distinctive invariant features from images. It has been successfully applied to a variety of computer vision problems based on feature matching including object recognition, pose estimation, image retrieval and many others. However, in real-world applications there is still a need for improvement of the algorithm‟s robustness with respect to the correct matching of SIFT features. An improvement of the original SIFT algorithm providing more reliable feature matching for the purpose of object recognition is proposed. The main idea is to divide the features extracted from both the test and the model object image into several sub-collections before they are matched. The features are divided in several sub-collections considering the features arising from different octaves.

The scale invariant feature transform (SIFT) algorithm, developed by Lowe is an algorithm for image features generation which are invariant to image translation, scaling, rotation and partially invariant to illumination changes and affine projection. Calculation of SIFT image features is performed through the four consecutive steps From the algorithm description given it is evident that in general, the SIFT-algorithm can be understood as a local image operator which takes an input image and transforms it into a collection of local features. To use the SIFT operator for object recognition purposes, it is applied on two object images.

- **BPN Face Recognition**

A new method, face recognition based on back propagation neural network, is presented. The proposed method extracts feature from face image with differential projection and geometrical features into eigenvector which is classified by back propagation neural network. Besides our method, the principal component analysis (PCA)-based method, the linear discriminated analysis (LDA) based method and the Markov Random Fields (MRF)-based methods were also tested for comparisons. The experimental results on ORL face database show that the proposed method achieves an average recognition accuracy of over 98% by using only 13 features. Moreover, the recognition accuracy is enhanced effectively, and the computational complexity and feature dimensions are reduced greatly.

Human beings have good recognition capabilities of faces and complex patterns and anything cannot affect this capability. This ability is quite robust, in spite of great changes in the visual stimulus due to facial expression, masking conditions, aging, and mismanagements such as whiskers, changes in hairdo or spectacles. The main reason for this is the high degree of interconnectivity, acquisition skills, adaptive quality, and abstraction capabilities of the human nervous system. There is various highly correlated biological neurons inhuman brain which can outperform super computers in certain specific tasks. Even a small child can perfectly and completely identify a human face, but it is a difficult task for the computer. Therefore, the main objective is to design such systems which can compete with what a small child does and thus making computers as lively as humans can. Face image is a biometrics physical feature which is used to verify the identity of people. The main components involved in the face image space include mouth, nose and eyes.

Back Propagation Neural Network (BPNN) is a multilayered and feed forward Neural Network. BPNN contains input layer, with one or many hidden layers that are being followed by the output layer. The layers contain identical computing neurons associated such that the input of every neuron in the next layer receives the signal from the output neuron in the previous layer. The input layer of the network serves as the signal receptor whereas the output layer passes the result out from the network

Face detection required that firstly noise should be reduced from the image so that better and more accurate results can be achieved. So, S. Adebayo Daramola proposed a system with four stages: face detection, pre-processing, principle component analysis (PCA) and classification. Firstly, database with images in different poses was made, then image was normalized and noise was removed from the image. After that, the Eigen faces were calculated from the training set. Then Eigen values were calculated using the PCA approach and then largest Eigen values were found comparing training set images and Eigen faces.

Hardware and Software Specification

Hardware Requirements

Hard Disk	:	40GB and Above
RAM	:	512MB and Above
Processor	:	Pentium III and Above

Software Requirements

Windows Operating System XP and Above
MATLAB 7.6.0(R2008)

Software Environment - **Matlab**

- High-level language for technical computing.
- Interpreter (allows fast prototyping).
- Scripting Programming Language.
- Just-In-Time Compilation (Since Mat lab 6.5).
- Integration of C, C++, Java-Code.
- Object-Oriented Programming,
- Mat lab compiler to build executable or a shared library using" Mat lab compiler runtime" in order to build applications which run without Mat lab.
- Matlab7.6.0 (R2008a)

VI. IMPLEMENTATION AND RESULTS

Implementation

First given Input short videos are converted into frames. . Then information less frames (Mean of Input frame<15) are eliminated. After resizing of each frame is done.Then all frames are merged into a single video for video categorization. Here different videos with different resolutions are taken. Here two videos are taken as shown in figure 4.1 and all the input videos are to be converted into frames Here the input videos which are taken as input should be in .avi format if not then convert all the different input videos into that format using total video converter software.

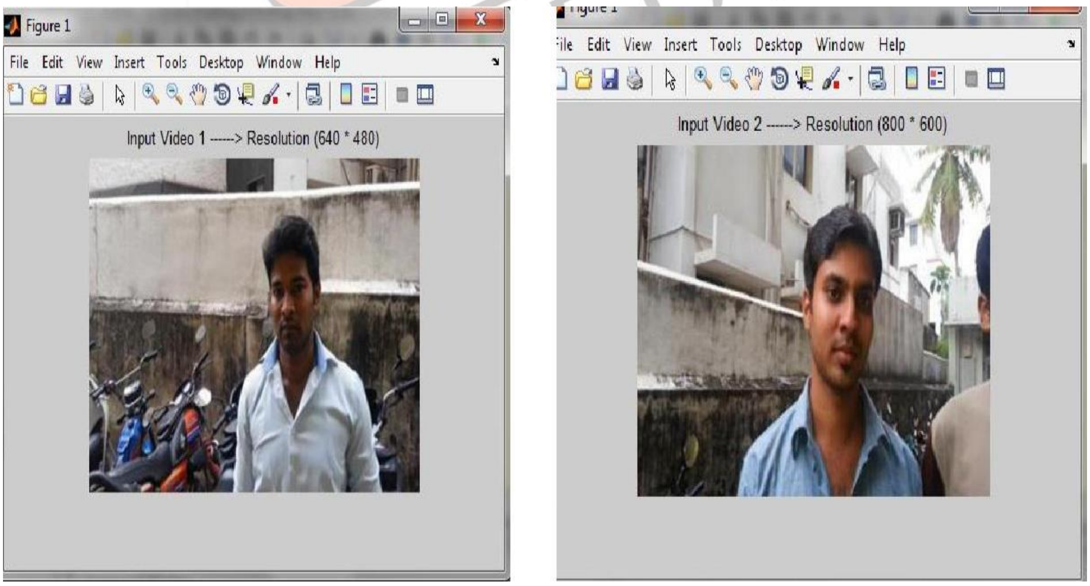
Figure 1
File Edit View Insert Tools Desktop Window Help
Input Video 1 -----> Resolution (640 * 480)
Figure 1
File Edit View Insert Tools Desktop Window Help
Input Video 2 -----> Resolution (800 * 600)

Fig 4.1: Input Videos 1,2

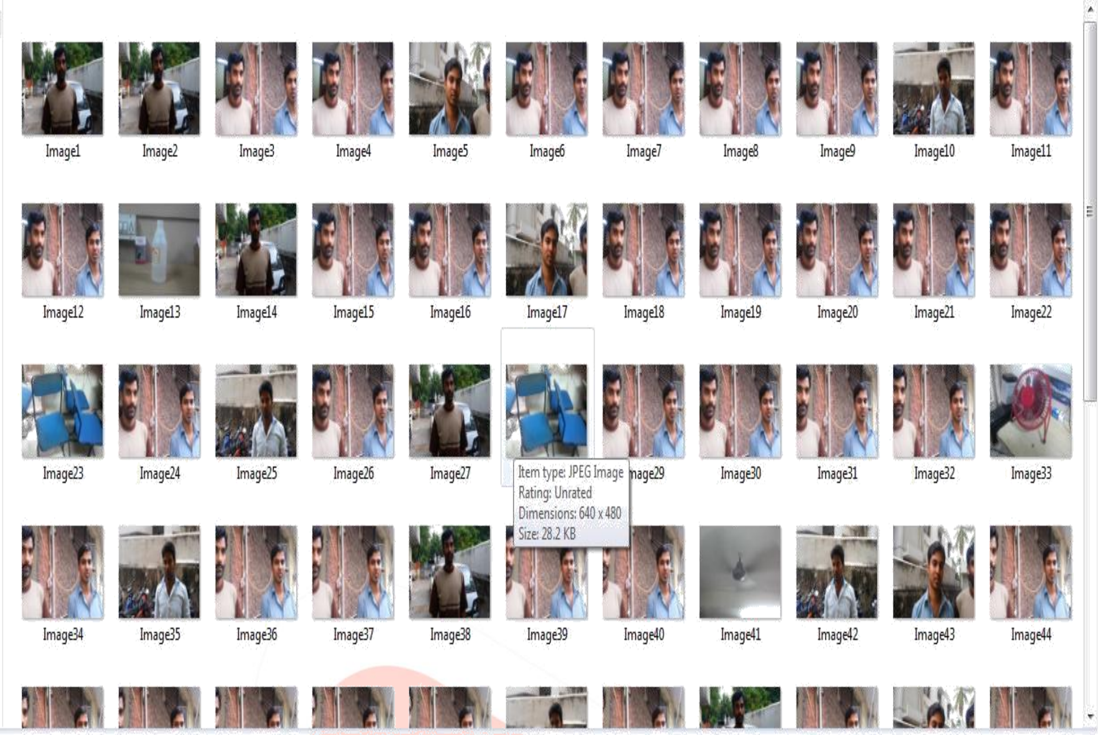

Fig 4.2: Input Video Frames

Figure 4.2 specifies that all the input videos which are taken can be converted into frames and here the frames which are converted have equal resolution. Input videos with different resolutions are taken after that convert all the videos into frames by using Mat lab syntax and all the frames are resized to equal resolution and finally merge that all the videos into video. This merged video having equal resolution In this way all the input videos are converted into frames after that all the frames are merged into a video.

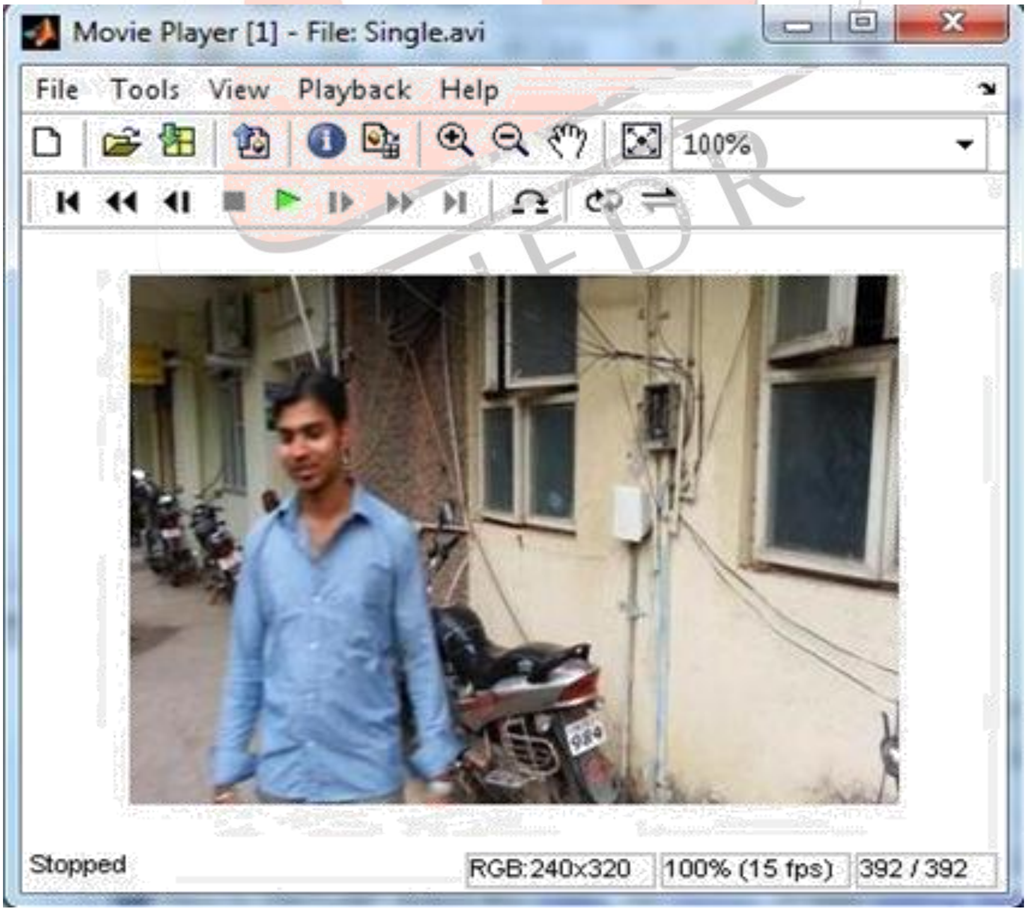

Fig 4.3: Merged Video

After converting all the videos into frames again all the frames are merged that is merged video and in this video all the frames have equal resolution. this video is as shown in fig 4.3,is taken as input video for further process. After this process we will apply viola Jones algorithm for face detection. Videos are categorized by using transition clues like human, object. Then human clue for first categorization is taken by using Viola-Jones algorithm, if faces are not detected in frames those frames are separated into another process for object matching clue.

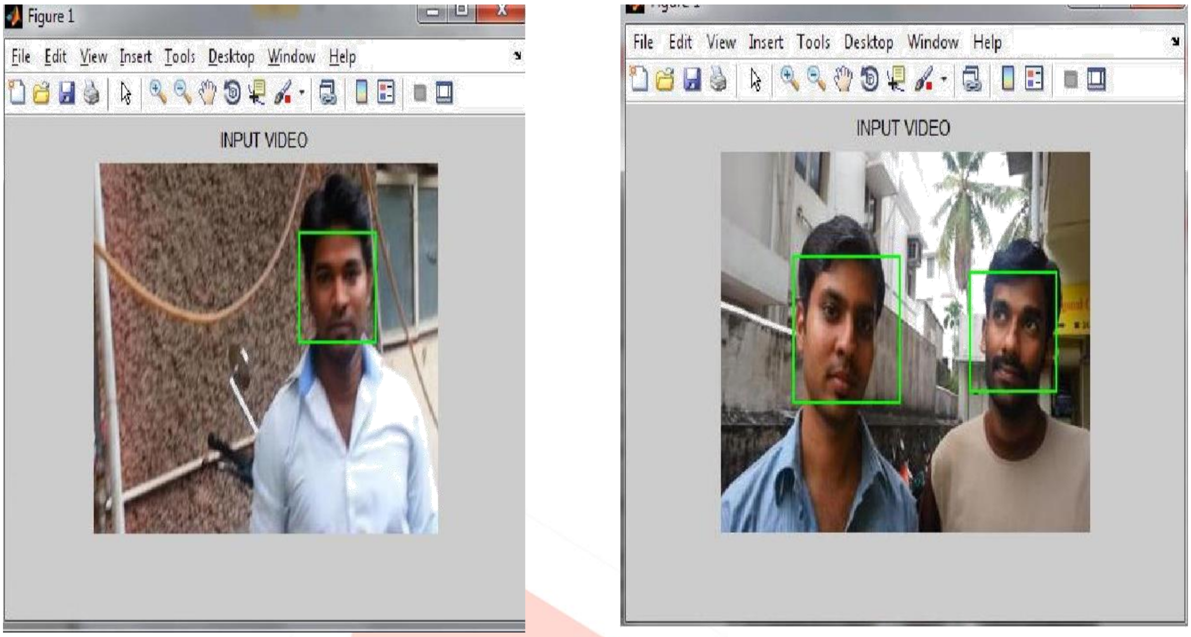
Figure 1
File Edit View Insert Tools Desktop Window Help
INPUT VIDEO
File Edit View Insert Tools Desktop Window Help
INPUT VIDEO

Fig 4.4: Applying Viola Jones Face Detection

The basic principle of the Viola-Jones algorithm is to scan a sub-window capable of detecting faces across a given input image as shown in fig 4.4. The standard image processing approach would be to rescale the input image to different sizes and then run the fixed size detector through these images. This approach turns out to be rather time consuming due to the calculation of the different size images. Contrary to the standard approach Viola-Jones rescale the detector instead of the input image and run the detector many times through the image each time with a different size.

First two directories are made like Human faces and Objects By using the command mkdir (make directory) in MATLAB. This human faces directory is used to detect human faces and object matching is used to detect objects. Here first take 1 rectangle with some length and width as per the human faces. By using that rectangle human faces are detected. All the frames containing human faces will be saved to human faces directory.

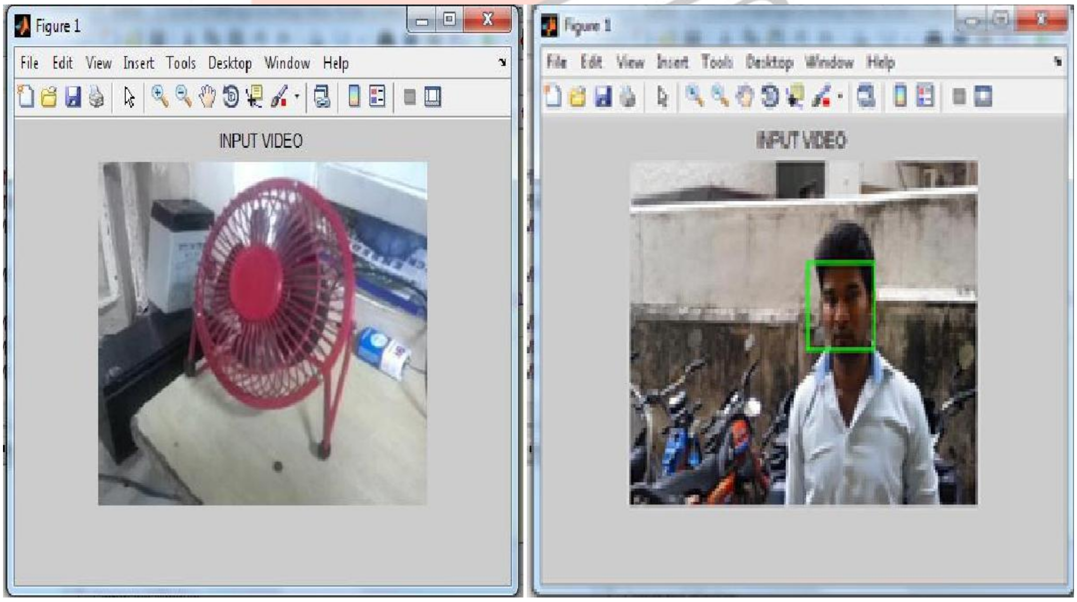
Figure 1
File Edit View Insert Tools Desktop Window Help
INPUT VIDEO
Figure 1
File Edit View Insert Tools Desktop Window Help
INPUT VIDEO

Fig 4.5: Separation of Both Human faces and Objects

The figure 4.5 shows that objects and human faces categorized separately. After applying the viola Jones algorithm. In this viola Jones algorithm by using that rectangle it will consider it will consider some part in humans that is face according to that given specifications. After finding human faces the remaining parts it will consider as object.

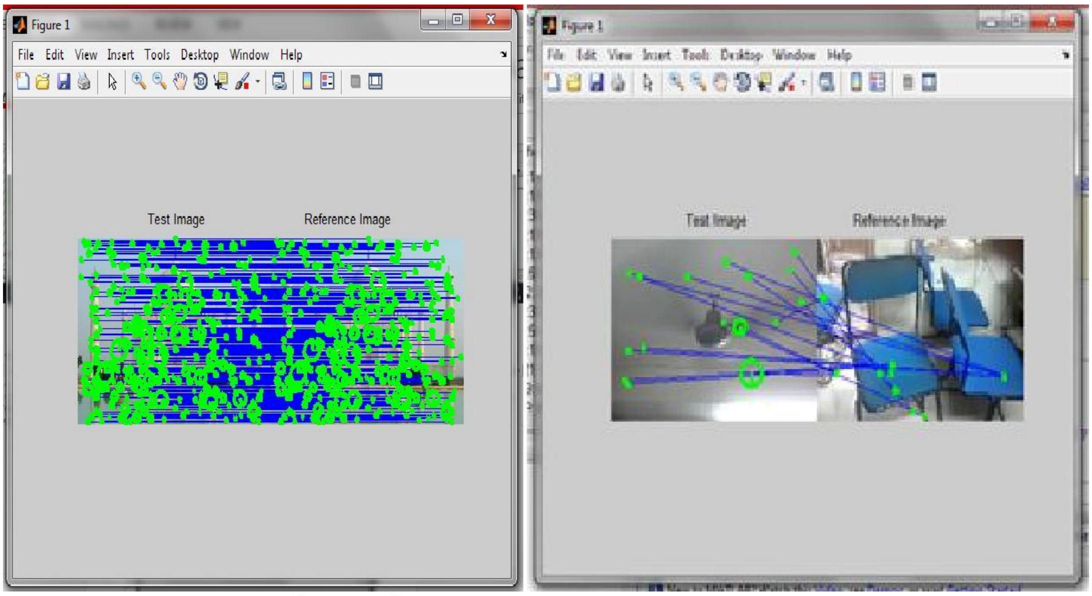
Figure 1
File Edit View Insert Tools Desktop Window Help
Test Image
Reference Image
Figure 1
File Edit View Insert Tools Desktop Window Help
Test Image
Reference Image

Fig 4.6: Comparison of Test and Reference Images Using SIFT Algorithm

In SIFT algorithm we will take all the objects which categorized previously. And store all the objects in one folder. From that one object is taken as reference object and remaining all will be taken as test objects. The object which we want as separate video that is reference object will compare with all the test objects if it matches means then all will take into 1 folder and will take as 1 video.

Fig 4.7: Particular Object Video

After applying SIFT algorithm the reference image will compare with test images all given and finally produce the particular object video as separately That video only shown above. This SIFT algorithm is mainly used for categorizing objects separately.

BPN Face Recognition

After applying viola Jones algorithm the human faces were separated and objects also separated. For that human faces we will apply BPN face recognition algorithm for making that human faces as one video and in that we will make some procedures like comparing reference face with all the input faces and if matching occurs means then it will take that matched person video as a separate video. By using this BPN face recognition we can categorize human faces separately in an efficient manner compared with the previous methods.

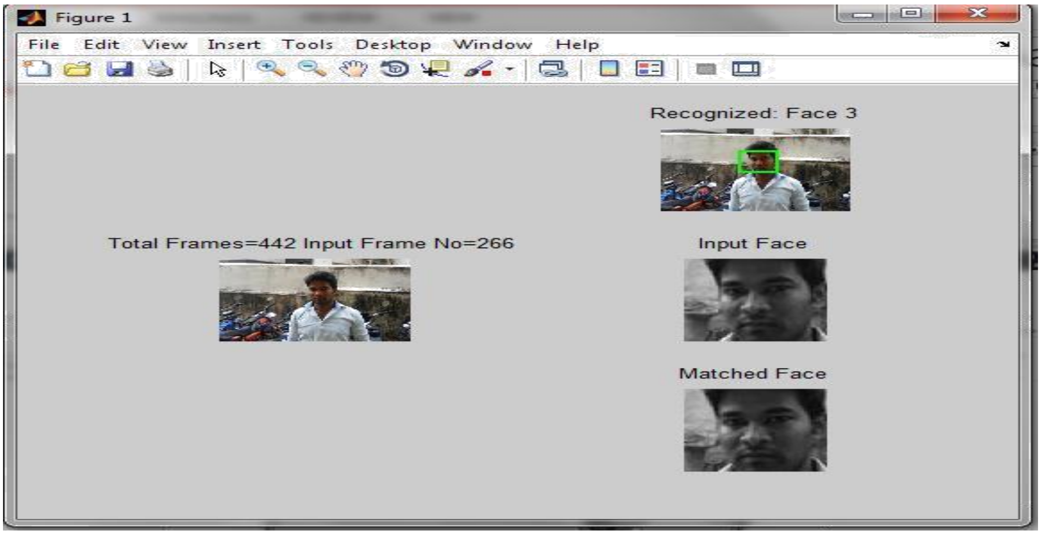
Figure 1
File Edit View Insert Tools Desktop Window Help
Recognized: Face 3
Total Frames=442 Input Frame No=266
Input Face
Matched Face

Fig 4.8: BPN Face Recognition

After applying BPN face recognition as shown in fig 4.8, the videos will be separated according to our requirement .That means if we place any reference face in the database according with that face we will get that particular face as 1 video as shown in the fig 4.9.

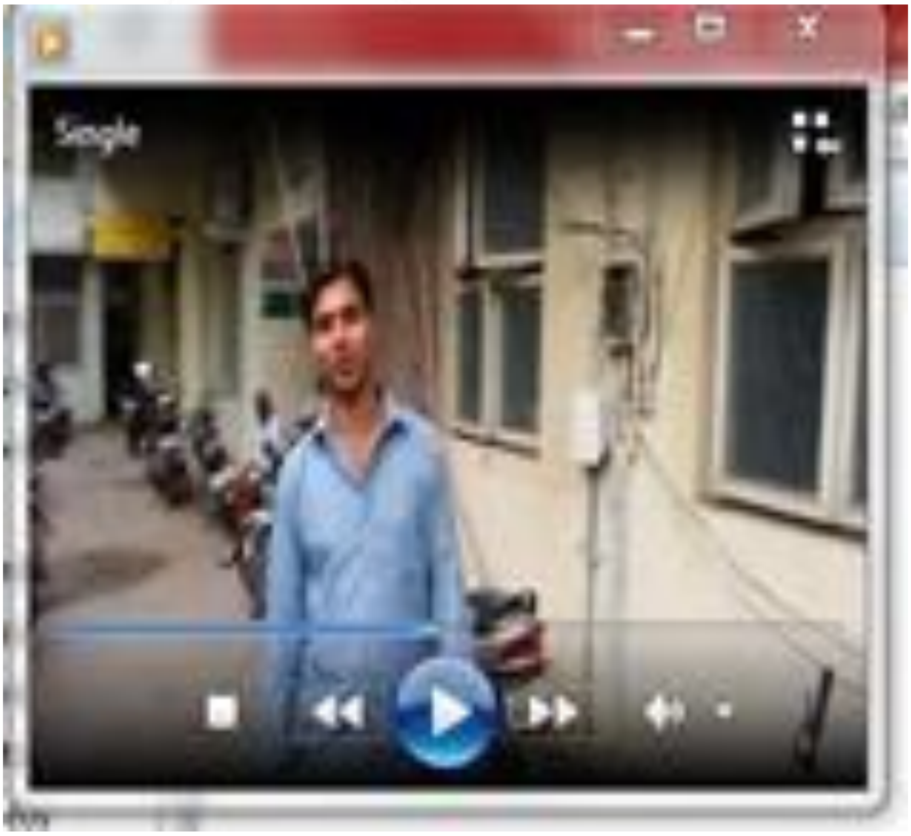

Fig 4.9: Output Video particular person.

Result

With the help of mat lab software all the given input videos are converted into frames, then they are preprocessed. After pre processing all the frames are converted into single video. Now this video is checked frame by frame, to separate frames containing human faces from other objects.

Fig 4.10: Snapshot of Result

After separating the frames, each person and object frames are detected and separated accordingly. Thus the separated frames are converted into videos, which is our desired result.

VII. CONCLUSION

In this project, we proposed "Video Puzzle", which is an integrated system for 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 a one-shot presentation using them. It can facilitate family album management and web video categorization.

REFERENCES

[1] Ahanger.G, "Automatic composition techniques for video production,"IEEE Trans. Knowl. Data Eng., vol. 10, no. 6, pp. 967–987,Nov. 1998.
[2] Barnes.C, Goldman.D, Shechtman.E, and . Finkelstein .A,"Video tape stries with continuous temporal zoom," in Proc. SIGGRAPH, 2010.
[3] Bennett.E, "Computational time-lapse video," ACM Trans. Graph.,vol. 26, no. 102, Jul. 2007.
[4] Bhat K.S,, Seitz S.M.,,Hodgins J.K,Khosla P.K, Bhat, K.S, .Seitz S.M, Hodgins J.K,, and Khosla P.K., "Flow-based video synthesisand editing," ACM Trans. Graph., vol. 23, no. 3, pp. 360–363, Aug.2004.
[5] Calic.J, Gibson .D, and Campbell .N, "Efficient layout of comic-likevideo summaries," IEEE Trans. Circuits Syst. Video Technol., vol. 17,no. 7, pp. 931–936, Jul. 2007.
[6] Caspi.Y, Axelrod .A, Matsushita .Y, and Gamliel .A "Dynamic stillsand clip trailers," Visual Comput., vol. 22, no. 9, pp. 642–652, Sep.2006.
[7] Chiu.P, Girgensohn.A, and Liu.Q, "Stained-glass visualization forhighly condensed video summaries," in Proc. ICME, 2004.
[8] Correa.C, "Dynamic video narratives," ACM Trans. Graph., vol. 29, no. 4, Jul. 2010
[9] Felzenszwalb P.F,, Girshick R.B,. McAllester.D, and Ramanan.D, "Object detection with discriminatively traine part-based models," IEEE Trans. Pattern Anal.Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[10] Nikolaidis C.C, "Video shot detection and condensed representation. A review," IEEE Signal Process. Mag., vol. 23, no. 2, pp. 28–37, Mar. 2006.
[11] Philbin O.C., "Near duplicate image detection: min-hash and tf-idf ,"in Proc. BMVC, 2008