International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 3 595 – 600

Extraction of Exclusive Video Content from One-Shot Video

A.Chithra
Student: M.E Applied Electronics
Velammal Engineering College
Chennai, Tamil Nadu, India
chithraa27@gmail.com

M.C.Shanker
Assistant Professor
Velammal Engineering College
Chennai, Tamil Nadu, India
rekshan@rediffmail.com

Abstract: With the popularity of personal digital devices, the amount of home video data is growing explosively. 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. This video composition system is used to generate aesthetically enhanced long-shot videos from short video clips. Our proposed system is to extract the video contents about a specific topic and compose them into a virtual one-shot presentation. All input short video clips are pre-processed and converted as one-shot video. Video frames are detected and categorized by using transition clues like human, object. Human and object frames are separated by implementing a face detection algorithm for the input one-shot video. Viola Jones face detection algorithm is used for separating human and object frames. There are three ingredients in this algorithm, working in concert to enable a fast and accurate detection. The integral image for feature computation, adaboost for feature selection and an attentional cascade for efficient computational resource allocation. Objects are then categorized using SIFT (Scale Invariant Feature Transform) and SURF (Speed Up Robust Features) algorithm.

Keywords:Single shot, video composition, face detection, integral image, adaboost, attentional cascade, SIFT, SURF

*****

I.INTRODUCTION

Video is an electronic medium for the recording, copying and broadcasting of moving visual images. Analog video is represented as a continuous time varying signal. Digital video is represented as a sequence of digital images. Composition refers to the organization of pictorial elements in a frame. Every image should have a single story to tell. The purpose of composition is to direct your viewer's eye to the central point or story in the scene. One-shot videos or long-shot video, also known as long-take video means a single shot that is with relatively longer duration. Long shot has been widely used in the professional film industry. 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 difficult even for professionals. In comparison with still image editing and composition, content based video editing and composition faces the additional challenges of maintaining the spatial-temporal consistency with respect to geometry.

Face detection is a computer technology that determines the locations and sizes of human faces in arbitrary digital images. It detects facial features and ignores anything else, such as buildings, trees and bodies. Face detection can be regarded as a specific case of object-class detection. In object-class detection, the task is to find the locations and sizes of all objects in an image that belong to a given class. Face detection can be regarded as a more general case of face localization. In face localization, the task is to find the locations and sizes of a known number of faces. In face detection, one does not have this additional information.

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Object recognition in computer vision is the task of finding and identifying objects in an image or video sequence. Humans recognize a multitude of objects in images with little effort, despite the fact that the image of the objects may vary somewhat in different viewpoints, in many different sizes / scale or even when they are translated or rotated. Scale-invariant feature transform (or SIFT) is an algorithm in computer vision to detect and describe local features in images. Applications include object recognition, robotic mapping , 3D modelling, gesture recognition, video tracking. SURF (Speed Up Robust Features) is a robust local feature detector, that can be used in computer vision tasks like object recognition or 3D reconstruction. It is partly inspired by the SIFT descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT. SURF is based on sums of 2D Haar wavelet responses and makes an efficient use of integral images.

II.RELATED WORKS

Cotsaces et al [2] described the techniques for organizing the video data into more compact forms or extract semantically meaningful information. The advances in digital video technology and the ever-increasing availability of computing resources have resulted in the last few years in

 ISSN: 2321-8169
595 – 600

an explosion of digital video data, especially on the Internet. However, the increasing availability of digital video has not been accompanied by an increase in its accessibility. This is due to the nature of video data, which is unsuitable for traditional forms of data access, indexing, search, and retrieval. The basic tasks used were shot boundary detection and condensed video representation.Xian-Sheng Hua et al [3] proposed this system which automatically selects suitable or desirable highlight segments from a set of raw home videos and aligns them with a given piece of incidental music to create an edited video segment to a desired length based on the content of the video and incidental music. They developed an approach for extracting temporal structure and determining the importance of a video segment in order to facilitate the selection of highlight segments.According to Gulrukh Ahanger [4] and Thomas, video production involves the process of capturing, editing and composing video segments. Composition must yield a coherent presentation of an event. This process can be automated if appropriate domain specific metadata are associated with video segments. This paper is proposed by identifying some fundamental attributes. Temporal continuity that characterizes the sequencing of segments in time. Themic continuity that ensures the smooth flow of conveyed information between consecutive segments and structural continuity that should have domain dependent structure.

Pedro F. Felzenszwalb et al [5] described an object detection system based on mixtures of multiscale deformable part models. This system is able to represent highly variable object classes and achieves state-of-the-art results in the PASCAL object detection challenges. Shatin, N.T [6] proposed this work by first, analyzing the structure of the video, and the boundaries of video scenes, and then calculating each scene's skimming length based on its structure and content entropy. Second, defining a spatial-temporal dissimilarity functions between video shots and model each video scene as a graph, then finding each scene's optimal skimming in the graph with dynamic programming. Finally, the whole video's skimming is obtained by concatenating the skimming's of the scenes. Experimental results show that this approach preserves the scene level structure and ensures balanced coverage of the major contents of the original video. Chang Huang et al [7] proposed this work for rotation invariant multiview face detection (MVFD) that aims to detect faces with arbitrary rotation-in-plane (RIP) and rotation-off-plane (ROP) angles in still images or video sequences. Paul Voila [8] and Michael Jones described a face detection framework that is capable of processing images extremely rapidlywhile achieving high detection rates.

III.PROPOSED METHODOLOGY

The video composition for a particular scene or object is described in the Fig.1 which explains the extraction of specific video content from a large number of short input video clippings.

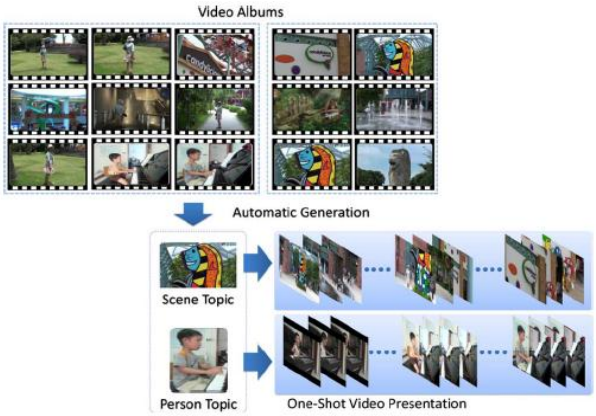


Fig 1 Overview of the composition of one-shot video from a selected scene or person

The input short video clips are converted to their respective frames and saved to separate folders corresponding to each input video. Resize all the frames to a particular specified resolution Meaningless frames are deleted by calculating the mean value of every frame and by setting a threshold for mean value. Followed by this, the resized useful frames are moved to a separate folder. These frames are then composed to form a one-shot video. The system overview is given in Fig.2

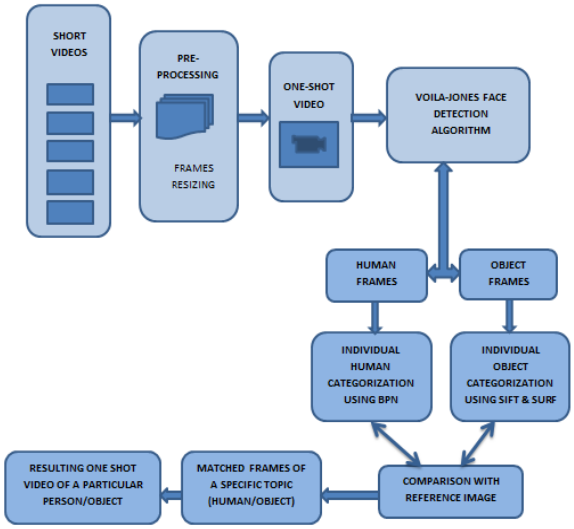


Fig.2. System Overview

A face detection algorithm is used for this composed video to detect human faces. Human and object frames can be detected and separated in two different folders by this method. Further categorization of a particular person or object can be accomplished by SIFT algorithm. SURF algorithm can be used for categorization in order to achieve a better matching score. Thus video contents about a specific topic can be extracted which is again composed to a one-shot presentation.

Video summarization methods attempt to abstract the main occurrences, scenes, or objects in a clip in order to provide an easily interpreted synopsis. Due to the advances in digital content distribution digital video recorders, this digital content can be easily recorded. However, the user may NOT have sufficient time to watch the entire video or the whole of video content may not be of interest to the user. In such cases, the user may just want to view the summary of the video instead of watching the whole video. Thus, the summary should be such that it should convey as much information about the occurrence of various incidents in the video. Also, the method should be very general so that it can work with the videos of a variety of genre.

IV.FACE DETECTION AND RECOGNITION

Face detection is a computer technology that determines the locations and sizes of human faces in arbitrary digital images. It detects facial features and ignores anything else, such as buildings, trees and bodies. In our proposed system Viola-Jones algorithm is used for efficient and accurate face detection in videos. A face detector has to tell whether an image of arbitrary size contains a human face and if so, where it is. One natural framework for considering this problem is that of binary classification, in which a classifier is constructed to minimize the misclassification risk. Since no objective distribution can describe the actual prior probability for a given image to have a face, the algorithm must minimize both the false negative and false positive rates in order to achieve an acceptable performance. This task requires an accurate numerical description of what sets human faces apart from other objects. It turns out that these characteristics can be extracted with a remarkable committee learning algorithm called Adaboost, which relies on a committee of weak classifiers to form a strong one through a voting mechanism.

A. VIOLA-JONES FACE DETECTION ALGORITHM

Training is slow but detection is very fast and accurate. There are three ingredients working in concert to enable a fast and accurate detection: the integral image for feature computation, Adaboost for feature selection and an attentional cascade for efficient computational resource allocation.

1. Integral Image

Our face detection procedure classifies images based on the value of simple features. There are many motivations for using features rather than the pixels directly. The most common reason is that features can act to en-code ad-hoc domain knowledge that is difficult to learn using a finite quantity of training data.The feature-based system operates much faster than a pixel-based system. The simple features used are reminiscent of Haar basis functions. More specifically, we use three kinds of features. The value of a *two-rectangle feature* is the difference between the sum of the pixels within two rectangular regions. The regions have the same size and shape and are horizontally or vertically adjacent as given in Fig.3. A *three-rectangle feature* computes the sum within two outside rectangles subtracted from the sum in a centre rectangle. Finally a *four-rectangle feature* computes the difference between diagonal pairs ofrectangle

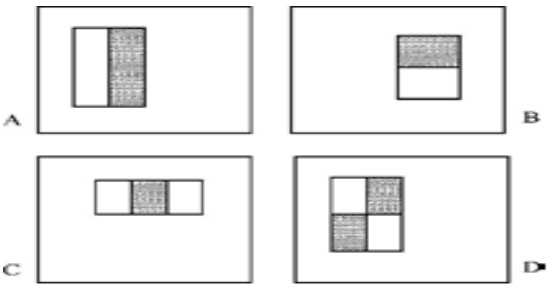


Fig.3.Example rectangle features shown relative to the enclosing detection window.

The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the grey rectangles. Two-rectangle features are shown in (A) and (B). Figure(C) shows a three-rectangle feature, and (D) a four-rectangle feature.

2. Adaboost for feature selection

For the task of face detection, the initial rectangle features selected by AdaBoost are meaningful and easily interpreted. The first feature selected seems to focus on the property that the region of the eyes is often darker than the region of the nose and cheeks as in Fig 4

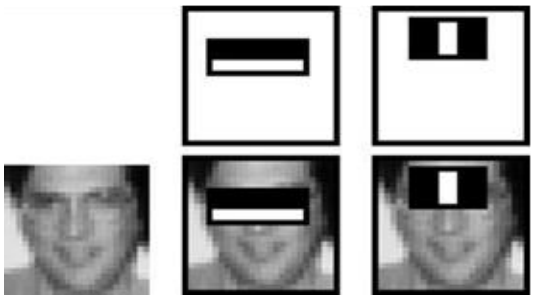


Fig.4. Adaboost Features

The first and second features selected by AdaBoost. Thetwo features are shown in the top row and then overlayed on a typical training face in the bottom row. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. The feature capitalizes on the observation that the eye region is often darker than the cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose.

Boosting Algorithm:

- Collect as many positive and negative samples
- Calculate integral image using rectangular features
- Initialize and updates the weights for the formed integral image
- Select the best weak classifier with respect to the weighted error

- Final strong classifier is obtained by cascading all the stages.

The key advantage of AdaBoost as a feature selection mechanism, over competitors such as the wrapper method, is the speed of learning. This feature is relatively large in comparison with the detection sub-window, and should be somewhat insensitive to size and location of the face. The second feature selected relies on the property that the eyes are darker than the bridge of the nose.

3. Attentional Cascade

This section describes an algorithm for constructing a cascade of classifiers which achieves increased detection performance while radically reducing computation time. The key insight is that smaller, and therefore more efficient, boosted classifiers can be constructed which reject many of the negative sub-windows while detecting almost all positive instances. Stages in the cascade are constructed by training classifiers using AdaBoost. Starting with a two-feature strong classifier, an effective face filter can be obtained by adjusting the strong classifier threshold to minimize false negatives. The overall form of the detection process is that of a degenerate decision tree, what we call a "cascade" see Fig. 5.

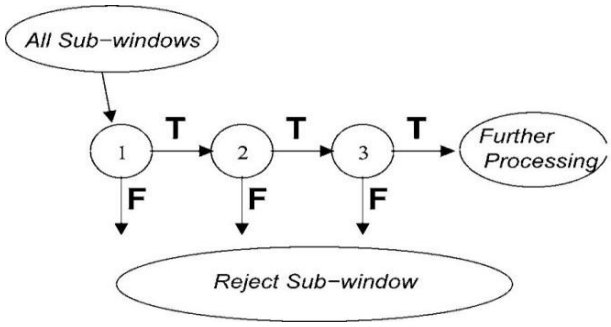


Fig.5. Schematic depiction of the detection cascade

A positive result from the first classifier triggers the evaluation of a second classifier which has also been adjusted to achieve very high detection rates. A positive result from the second classifier triggers a third classifier, and so on. A negative outcome at any point leads to the immediate rejection of the sub-window. The structure of the cascade reflects the fact that within any single image an overwhelming majority of sub-windows are negative.

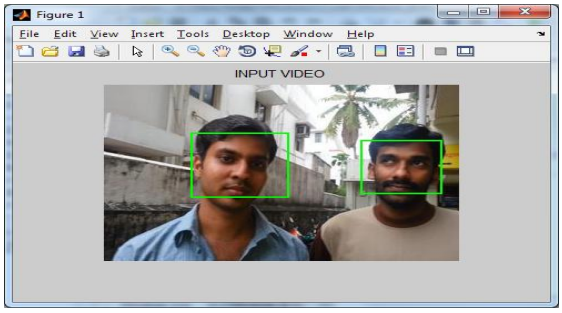


Fig.6. Face Detection using Viola Jones Algorithm

B. FACE RECOGNITION USING PCA BASED FEATURE EXTRACTION

Feature extraction of eigen values can be done by Principal Component Analysis (PCA**).** The basic concept of PCA is to find meaningful patterns or features in the input data without an external assistance.

Algorithm:

Step 1: Creation of Database

- Coordinate a set of face images
- Reshaping of the2D images of the training database into 1D column vectors
- Construct a 2 D matrix from 1D image vector

Step 2: Eigen values calculation

- Determine the most discriminating features between images of faces
- Compute mean of the face images in training database and calculate the deviation of each image from mean image
- Calculate the eigen values of the covariance matrix of the training database.

Step 3: Recognition of Image

- Project the centred image into face space and extract PCA features from the test image
- Calculate Euclidean distances between the projected test image and the projection of all centred training images.

V.OBJECT DETECTION AND CATEGORIZATION

Image matching is a fundamental aspect of many problems in computer vision, including object or scene recognition, and motion tracking.

A. Scale Invariant Feature Transform

This algorithm describes image features that have many properties that make them suitable for matching differing images of an object or scene. The features are invariant to image scaling and rotation. Large numbers of features can be extracted from typical images with efficient algorithms. The cost of extracting these features is minimized by taking a cascade filtering approach, in which the more expensive operations are applied only at locations that pass an initial test. Following are the major stages of computation used to generate the set of image features:

- Scale-space extrema detection
- Keypoint localization
- Orientation assignment
- Keypoint descriptor

The first stage of computation searches over all scalesand image locations. It is implemented efficiently by using a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation. At each candidate location, a detailed model is fit to determinelocation and scale. Keypoints are selected based

ISSN: 2321-8169
595 – 600

on measures of their stability. One or more orientations are assigned to each keypoint location based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations. Finally, the local image gradients are measured at the selected scalein the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

This approach has been named the Scale Invariant Feature Transform (SIFT), as it transforms image data into scale-invariant coordinates relative to local features. Generally, the high dimensionality of the descriptor is a drawback of SIFT at the matching step. For on-line applications relying only on a regular PC, each one of the three steps (detection, description, matching) has to be fast. Hence SURF algorithm can replace SIFT.

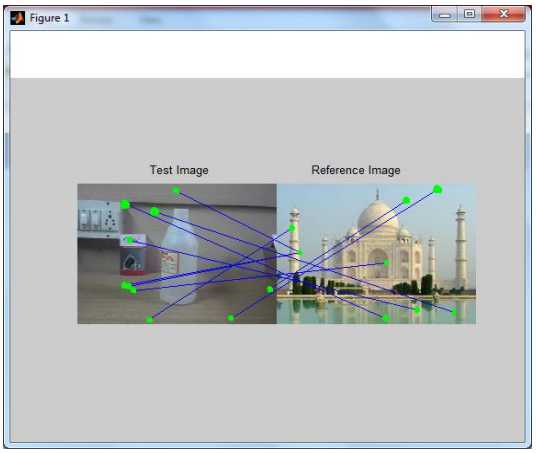


Fig 7.Keypoints matching between test and reference using SIFT

B. Speed Up Robust Features

Speeded Up Robust Features (SURF) approximates or even outperforms previously proposed schemes with respect to repeatability, distinctiveness and robustness. Our descriptor describes the distribution of the intensity content within the interest point neighborhood, similar to the gradient information extracted by SIFT and its variants. We build on the distribution of first order Haar wavelet responses in x and y direction rather than the gradient, exploit integral images for speed, and use only 64 dimensions. This reduces the time for feature computation and matching, and has proven to simultaneously increase the robustness. Furthermore, we present a new indexing step based on the sign of the Laplacian, which increases not only the robustness of the descriptor, but also the matching speed (by a factor of two in the best case). We refer to our detector-descriptor scheme as SURF Speeded-Up Robust Feature.

SURF algorithm is based on:

- Interest point detection
- Interest point localization
- Interest point description and matching

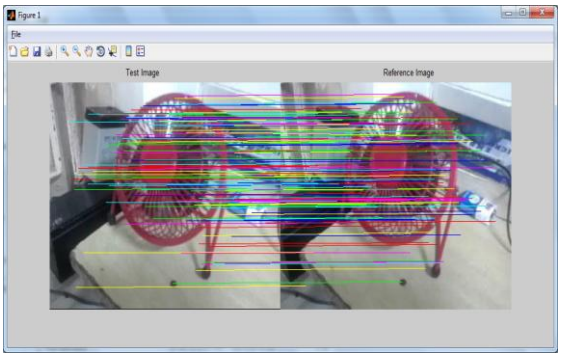
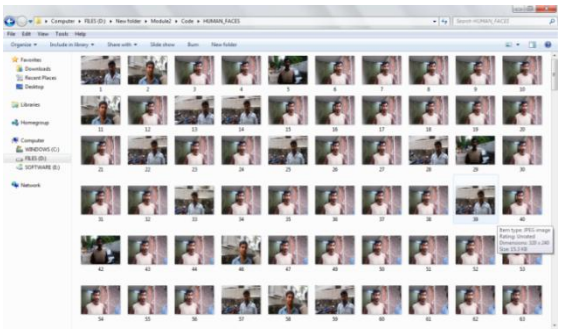
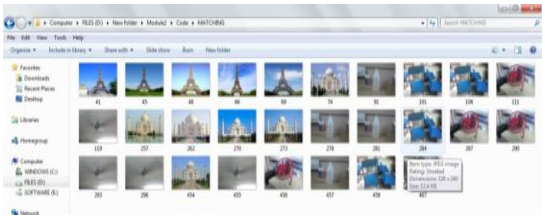


Fig.8 Interest points matching using SURF

VI. EXPERIMENTAL RESULTS

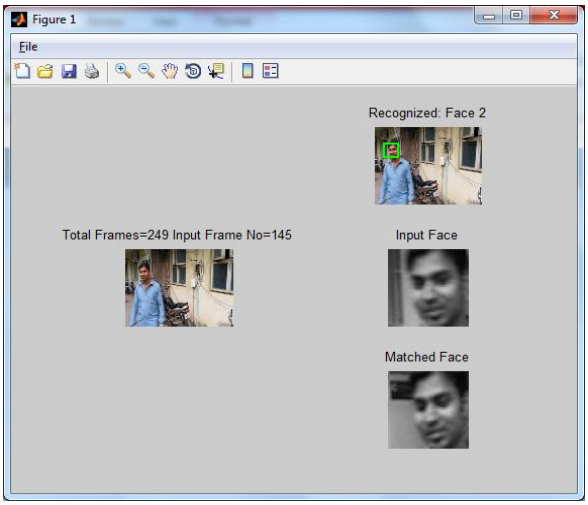


(a) Human Categorization



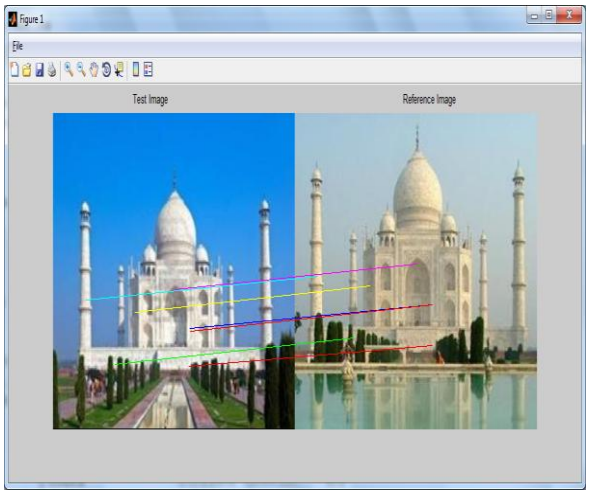
(b) Object Categorization

Fig.9.Application of face detection algorithm to distinguish human and object in two separate folders



(a) Matching of input recognized face

ISSN: 2321-8169
595 – 600



(b) Object Matching using keypoints

Fig.10. Matching of human and object with that of the reference image for individual categorization.

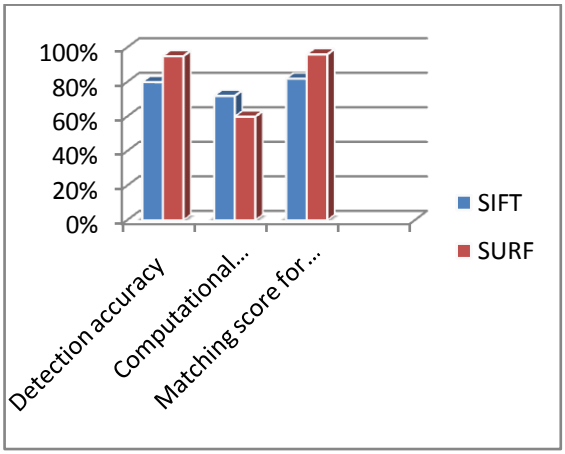


Fig.11. Comparison of SIFT and SURF algorithm for object detection

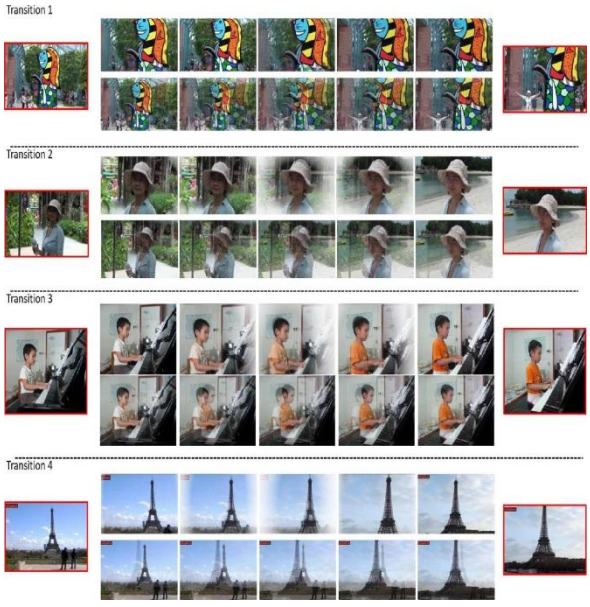


Fig.12. Resulting one-shot video composition of specific topic which is given as the reference image by the user

VII.CONCLUSION

Short videos are converted to their respective frames and moved to separate folders. All video frames are resized and mean values are calculated for every frame to eliminate the meaningless frames before composition. One-shot video is then composed from resized useful frames. First motive is to separate human and object frames which is implemented using various face detection algorithms. Part-based model provided some missing or false detections which is overcome using Voila Jones algorithm for face detection. This algorithm provided better results than previous. However the training is slow, but detection is very fast and hence suitable for video processing.

Followed by this object detection and categorization is implemented. 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 and can be used for various applications that involve video contents from surveillance cameras and when the user requires the video content about a specific person or object. This reduces the time for feature computation and matching, and has proven to simultaneously increase the robustness. As a future work this combined approach of human cum object detection and categorization can be implemented in real time videos.

VIII.REFERENCE

[1] Qiang Chen, Meng Wang "Video Puzzle Descriptive One-Shot Video Composition", IEEE Trans. On Multimedia, Vol. 15, No. 3, APRIL 2013

[2] C. C. Nikolaidis (2006), "Video shot detection and condensed representation-A review," IEEE Signal Process.Mag., vol. 23, no. 2, pp. 28–37.

[3] X.S.Hua, L.Lu, and H. J. Zhang (2004), "Optimization-based automated home video editing system," IEEE Trans. Circuits Syst. Video Technol.,vol. 14, no. 5, pp. 572–583.

[4] G. Ahanger, "Automatic composition techniques for video produc-tion," IEEE Trans. Knowl. Data Eng., vol. 10, no. 6, pp. 967–987, Nov. 1998.

[5] P.F.Felzenszwalb, R.B.Girshick, D.McAllester, and D.Ramanan (2010), "Object detection with discriminatively trained part-based models,"IEEE Trans. Pattern Anal.Mach.Intell., vol. 32, no. 9, pp. 1627–1645.

[6] S.Lu,I. King, andM.R. Lyu (2004), "Video summarization by video structure analysis and graph optimization," in Proc. ICME.

[7] C.Huang, H.Ai, Y.Li, and S.Lao (2007), "High-performance rotation invariant multiview face detection," IEEE Trans. Pattern Anal. Mach.Intell., vol. 29, no. 4, pp. 671–686.

[8] Paul Viola and Michael Jones (2004), "Robust Real-Time Face Detection", International Journal of Computer Vision, pp 137-154.

[9] D.Lowe (2004), "Distinctive image features from scale-invariant keypoints,"Int. J. Comput. Vision, vol. 60, no. 2, pp. 91–110.

[10] Yi-Qing Wang (2013), "An Analysis of Viola Jones Face Detection Algorithm", IPOL, ISSN 2105-1232