MULTIPLE PIPS DETECTION IN UNBOUNDED VIDEO STREAM

Chengdong Cui, Yao Zhao, Shikui Wei, Zhenfeng Zhu

Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China
Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, 100044, China
shkwei@bjtu.edu.cn

ABSTRACT

Picture-in-picture detection aims to detect the video clips that are embedded as a sub-window into a background video with similar semantic meaning, which plays an important role in various multimedia applications such as video copy detection and semantic relationship mining. While some PiP detection algorithms have been given in existing work, less effort has been made on the detection of multiple PiP clips in unbounded video streams. In this paper, we propose a successive-group prior of both PiP and non-PiP clips based on the statistics of single PiP detection accuracy. Using this prior, a new algorithm, which involves a series of geometrical and temporal constraints for handling both the spatial and temporal localizations, is proposed for detecting multiple PiP clips. The extensive experimental results show that the proposed algorithm can efficiently localize the spatial positions and precisely detect the starting and ending time stamps of multiple PiP clips.

Index Terms—Picture-in-Picture Detection, Video Copy Detection, Social Media, Cross-Media

1. INTRODUCTION

With the rapid development of computing and internet technology, the amount of digital videos is growing dramatically, which makes the data processing and analysis challenging in some multimedia applications, such as content based video copy detection, semantic relationship mining, event detection. Furthermore, the popularity of non-linear editing tools makes the problem more difficult and complex, since user can arbitrarily edit some original videos by employing some transformations like inserting, cropping, or flipping. As an important and tough video transformation, Picture-in-picture (PiP) is frequently used to combine and show the content from different video sources in a single video. Since the same frame contains the content from different sources, it is difficult to directly handle the video stream due to the limitation of image and video understanding techniques. In essence, the processing and analysis of video data are based on more compact feature representation rather than raw video data so as to simplify the processing step. Generally speaking, there are two kinds of feature representation schemes that are frequently used in multimedia community, i.e., global features and local features. For the global representation such as ordinal feature [1] [2] [3], since the extracting process is generally based on the information statistic of the whole frame or video clip, it cannot distinguish the content of embedding video from the background video. Therefore, global features cannot meet the requirement in some applications such as copy detection where we need to determine if the embedding video is derived from copyright protection video. For the local feature representation such as SIFT- and SURF-based schemes [4] [5], lots of local features are extracted from the whole frame, and embedding video can be separately represented by their local features. However, since we don't know if there are an embedding window beforehand, the video matching is still based on the whole frame, leading to wrong matching results. That is, whichever kind of features you choose, it will be inevitably impacted by PiP. Therefore, it is necessary to find an approach to separate embedding video windows from the background video before content analysis. In this paper, our main effort focuses on this important video preprocessing step, i.e., PiP detection.

The key challenge of PiP detection is how to precisely localize the embedding video clip. In the previous work [6] [7] [8], the main effort focuses only on spatial localization of embedding video. That is, the whole input video is generally treated as a detection unit, and the detection systems determine whether the video is PiP video and localize spatial region of the embedding video if it is. However, this kind of methods is based on an underlying assumption that all frames of the input video are either PiP frames or non-PiP frames. Hence, if the input video contains several discontinuous PiP clips, as shown in Figure 1, all these methods will fail to make correct decision. In addition, even if the whole video is PiP, these methods will make the computational complexity unacceptable when the input video is very long. Nevertheless, detecting multiple PiPs in unbounded video is useful in many application scenarios. For example, multiple PiPs frequently are inserted into news broadcast, where each PiP conveys one news story. In order to detect multiple PiPs in unbounded video streams, three

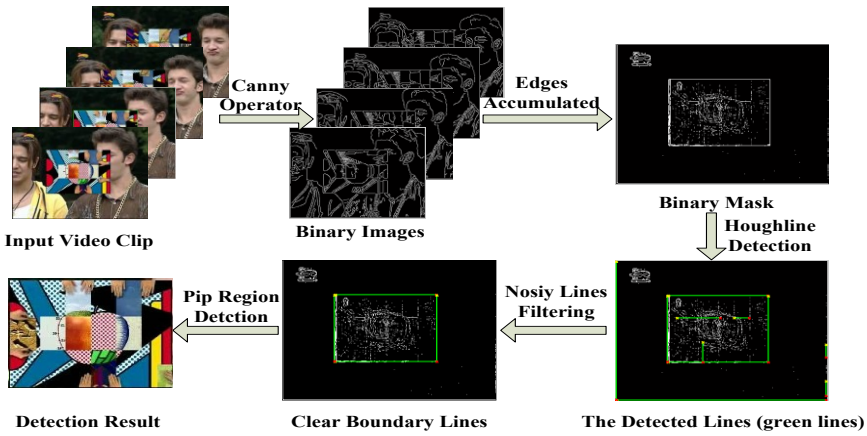

Fig. 2. Framework of the proposed single PiP detection scheme

Fig.1. Frame sequence of two discontinuous PiP clips with the sub-windows in the middle and in the upper right corner

key issues must be addressed: (1) How to determine whether there are some PiPs or not; (2) How to localize the spatial positions of embedding video windows; (3) How to distinguish PiP clips from non-PiP stream, i.e, temporal localization. To address these problems, we propose a multiple PiPs detection algorithm, called spatio-temporal PiP localizing algorithm. Specifically, we first design a single PiP detection scheme based on Hough transform and frame accumulation algorithm, which can effectively address single PiP detection problem for length-limited input videos. Then, we discover a successive-group prior of both overlapped PiP and non-PiP clips based on the statistics of single PiP detection accuracy. Finally, following the prior knowledge, we introduce a series of geometrical and temporal constraints into our PiP detection framework so as to handle both the spatial and temporal localizing problems. The experiments demonstrate that our scheme can efficiently localize the spatial positions and precisely detect the starting and ending time stamps of PiP clips.

2. SINGLE PIP LOCALIZING ALGORITHM

In this section, we first design an effective and efficient single PiP detection scheme, which is based on Hough transform and frame accumulation algorithm. The whole framework is illustrated in Figure 2. As shown, we sample key frames to represent the raw input video clip so as to avoid redundant and time-consuming processing procedure. In our experiment, we uniformly sample three frames per second. Given a video clip, the following four main steps are performed to detect the PiP clip.

Edge Detection: Extracting edge information of video frame is performed as a preprocessing step, since the region of embedding video can be treated as a kind of regular edges. In our scheme, we employ the Canny operator to perform the edge extracting process [9]. After extracting edge information, a sequence of binary edge images is obtained as shown in Figure 2.

Edge Accumulation: According to the assumption discussed above, the edges located at embedding video windows should be always retained across the whole detection unit. By accumulating all edges of all frames, the amplitudes located at PiP window boundary will be quite different to other positions, which will facilitate the localization of PiP windows. In this step, all the edge images in sequence are accumulated to a single edge image A. Furthermore, the accumulated image is mapped to a binary mask B by a threshold function which is defined as follows:

$$B(i,j)=\begin{cases}1, A(i,j)>k\\ 0, A(i,j)\le k\end{cases} \quad (1)$$

where $A(i,j)$ and $B(i,j)$ denote the value of A and B *at point* (i,j), respectively; k is a fixed threshold value and is set to $0.2\times l$, l denotes the length of detection unit. It means that only the edges that frequently occur can be reserved.

Hough Transform based Line Detection: Line detection method based on Hough transform [10] is employed to remove noisy points and explicitly identify the straight lines in the binary edge image B. Meanwhile, some noisy lines such as short lines are filtered out, which results in a clear boundary image of lines.

Region Localization: Actually, there are still lots of noisy lines even after line filtering step in the practical scenario. Therefore, we introduce a new method to identify the position of embedding video window in this step, which is

based on a pair of orthogonal line segments. Here, the pair of orthogonal lines is denoted as $P\{H,V\}$, where H represents the horizontal line and V represents the vertical line. The two points of (x_1,y_1) and (x_2,y_2) mark the ends of line segment H, and (m_1,n_1), and (m_2,n_2) denotes the ends of line segment V. We identify a PiP window if the pair of orthogonal line segments meets both of the following conditions:

$$\begin{aligned} |x_i - m_j| &< t \\ |y_i - n_j| &< t \end{aligned} \tag{2}$$

where i=1or2, j=1or2, t is the allowed deviation at the position where the orthogonal line segments or their extension cross with each other. In our context, t is set to 10. In essence, the proposed method is designed by considering the following three geometry constraints [6]:

- Distance constraint: The parallel lines of a rectangle should not be too close or too far away.
- Location constraint: The vertical lines should be located between the two horizontal lines, and the horizontal lines should be located between the two vertical lines.
- Crossing constraint: The extensions of the two orthogonal lines of a rectangle cannot cross at the middle of each line.

Generally speaking, the positions of rectangles are different from each other because they are generated by different pairs of orthogonal lines. However, if the pairs of orthogonal lines belong to the same sub-window, there may be several rectangles with approximate position. In this case, we will only choose the rectangle with the biggest area as the final PiP region. By employing the abovementioned four steps, we precisely localize the PiP region if there is a PiP, otherwise nothing is reported.

3. MULTI-PIPS LOCALIZING ALGORITHM

As discussed above, the existing PiP detection algorithms fail to handle multi-PiPs case in unbounded video stream. To address the problem, we propose a spatio-temporal PiP localizing algorithm, which involves successive-group prior knowledge discovering and spatio-temporal joint localizing.

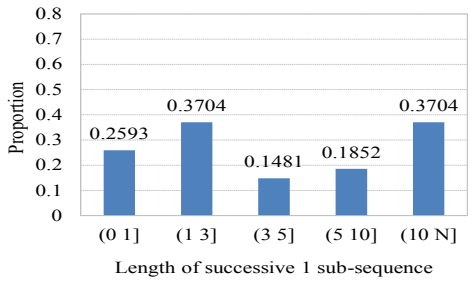

Fig. 3. The statistics on various lengths of successive 1 sub-sequences from PiP video streams.

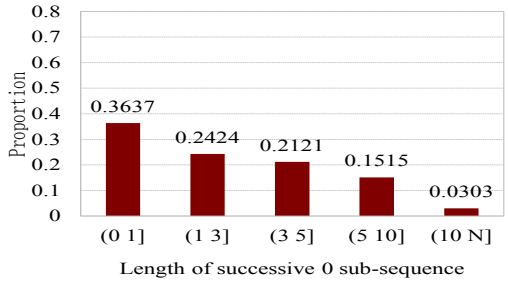

Fig. 4. The statistics on various lengths of successive 0 sub-sequences from PiP video streams.

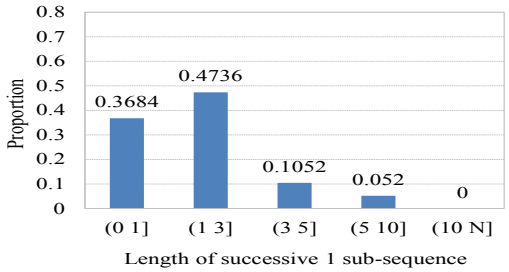

Fig .5. The statistics on various lengths of successive 1 sub-sequences from non-PiP video streams.

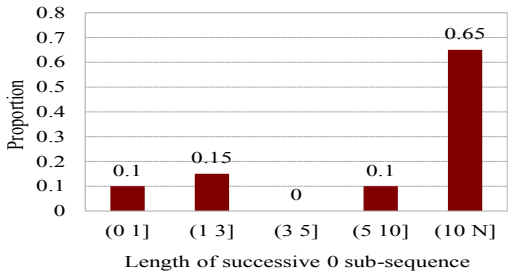

Fig .6. The statistics on various lengths of successive 0 sub-sequences from non-PiP video streams.

3.1. Successive-group prior discovering

In order to handle the unbounded video streams, we have to perform lots of single PiP detection processes by moving a length-fixed sliding window in the time axis. However, due to the accuracy limitation of detection algorithm, there are always some missed or false PiP clips when shifting the window in the stream, which will greatly affect the temporal localization precision. In this section, we attempt to discover some rules about the missed PiPs and false PiPs. In order to investigate the rules about missed PiPs, we select two long video streams with the same length, and one stream is embedded as a video sub-window into the other one. Then we perform the proposed single PiP detection algorithm on a clip within a length-fixed sliding window in video stream. If the algorithm asserts the clip to a PiP clip, it outputs 1, otherwise 0. By shifting the sliding windows with a fixed

step along the video stream, we can obtain a long binary sequence. We make some statistics on lengths of both successive 0 sub-sequences and successive 1 sub-sequences. As shown in Figures 3 and 4, most of the successive 0 sub-sequences are very short, while the successive 1 sub-sequences are much longer in PiP video streams. This means that we can treat a short successive 0 sub-sequence as missed PiPs and correct it to 1 sequence if the short successive 0 sub-sequence is located between two long successive 1 sub-sequences. Likewise, similar rules can be obtained on false PiPs from long video stream without any PiP clips. Figures 5 and 6 demonstrate the statistics, where the successive 1 sub-sequences means false PiP clips. We call these rules as successive-group prior.

3.2. The flexible temporal localizing scheme

The key issue for detecting multiple PiP clips in unbounded video streams is how to determine the temporal boundaries. Given an unbounded video stream, the scheme starts the detecting procedure by performing single PiP detection on a video clip within a fixed temporal window in the stream. By moving the window forward with a step, we can obtain a sequence of decisions, which is denoted as $I=\{F_1(t_1),F_2(t_2),...F_i(t_i),...|i\geq 1\}$. When the current detection unit i is asserted a PiP clip, F_i is set to 1, otherwise 0. t_i indicates the time stamp of current detection unit. To alleviate the effect of missed and false PiP clips, the decision sequence I is preprocessed by the following two steps, which is designed according to the successive-group prior above.

Firstly, if the current decision F_i is 1, and the following conditions

$$F_{i-1}=0 \text{ or } F_{i+1}=0 \tag{3}$$

are satisfied. Then the F_i is updated to 0, i.e, correcting the decision of false PiPs. After updating all the elements in I, a new decision sequence I_1 is obtained.

Secondly, if the current decision F_i in decision sequence I_1 is 0, and the following conditions

$$F_{i-1}=1 \text{ or } F_{i+1}=1 \tag{4}$$

are satisfied. Then the F_i is updated to 1, i.e, correcting the decision of missed PiPs. After updating all the elements in I_1, a new decision sequence I_2 is obtained.

Once we have obtained decision sequence I_2, the temporal boundaries of PiP clips can be decided. If F_i is 1, we set its time stamps t_i as the starting point of a PiP clip. If F_j is 1, and the following conditions

$$\begin{cases} F_{j-1}=1 \\ F_{j+1}=0 \end{cases} and \begin{cases} F_{j+2}=0 \\ F_{j+3}=0 \end{cases} \tag{5}$$

are satisfied. Then the t_j is treated as the ending point of a PiP clip.

Since a starting point s_i is always followed by an ending point e_i, we denotes $S(s_i,e_i)$ as time stamps of the asserted PiP clip in the video stream. As discussed above, missed PiP clips will split a desired clip into lots of short PiP clips. To address this problem, we introduce a flexible strategy to merge the possible split PiP clips, which is defined as follows:

- If $0<s_{i+1}-e_i\leq l$, The two PiP segments (s_i,e_i) and (s_{i+1},e_{i+1}) will be directly joined into one clip.
- If $l<s_{i+1}-e_i\leq 2l$, one more condition will be considered. The two PiP segments can be joined into one clip only if the spatial regions of the clips are the same.
- If $2l< s_{i+1}-e_i$, the two PiP segments will be directly regarded as two independent PiP clips.

where l indicates the length of the detection unit.

4. EXPERIMENT

4.1. Data set

The proposed spatio-temporal PiP detection algorithm is evaluated on the Sound & Vision dataset used in TRECVID 2008 search and high level feature extraction tasks. This dataset contains approximately 200h of videos including news magazine, science news, educational programming, and archival video. To evaluate the performance of multi-PiP clips detection, 25 video archives are selected randomly from the dataset as background video streams, and several short clips from different sources are treated as embedding videos. In order to clearly evaluate the miss rate and the false alarm rate, only 20 selected background video streams are individually embedded with two short embedding clips, and the other 5 background video streams are regarded as non-PiP video streams. To compare with existing single-PiP detection methods, another 20 video streams, which contain PiP frames across the whole video, are constructed from TRECVID dataset for evaluating single-PiP detection.

4.2. Evaluation criteria

In our experiments, two commonly used criteria, i.e., the miss rate and false alarm rate, are employed for evaluating the overall detection accuracy, which are defined as follows:

$$R_{Miss}=\frac{N_{rel}-N_{ret\_rel}}{N_{rel}} \tag{6}$$

$$R_{FA}=\frac{N_{ret}-N_{ret\_rel}}{N_{ret}} \tag{7}$$

where, N_{ret} is the total number of the returned results; N_{ret_rel} is the number of true positives in the returned results, and N_{rel} is the total number of true positives.

While both evaluation criteria can give a good measurement on the overall detection performance, they do

not take the temporal localization precision into account. Therefore, an additional evaluation criterion about temporal localization, i.e., overlap degree, are given as follows:

$$D(i) = \frac{overlap(P_i, P_{gi})}{Length\_Max(P_i, P_{gi})} \quad (8)$$

where $overlap(P_i,P_{gi})$ is the time span of the overlap between the asserted PiP clip P_i and its ground truth P_{gi}, $Length\_Max(P_i,P_{gi})$ is the time interval between the minimum starting point and the maximum ending point of two clips. Clearly, the bigger the overlap degree is, the better the temporal localization precision is.

To evaluate the relationship between detection accuracy and temporal localization precision, we define threshold detection accuracy as follows:

$$accuracy = \frac{N_T}{N_{rel}} \quad (9)$$

where N_T is the number of true positives whose overlap degree is higher than a threshold T.

Due to imperfection in either the video data or the PiP detection algorithm, it is possible to split one desired PiP clip multiple sub-segments. Therefore, it is necessary to define a criterion for analyzing the scatter degree, which is denoted as follows:

$$R_{Dis} = \frac{N - N_{rel}}{N_{rel}} \quad (10)$$

where N is the total number of asserted PiP clips that are true positives.

4.3. Evaluation on temporal localization

In this section, we evaluate the temporal localization precision of the proposed method. As discussed above, it is possible that one desired PiP clip is divided into several discontinuous sub-PiP clips. To handle this case, here, we only evaluate two longest PiP clips detected from each testing video stream, since there are at most two PiP clips for each video stream. In fact, the number of the PiP clips which are divided into multiple sub-clips is small. Figure 7 shows the number of asserted PiP clips and the number of groundtruth. As shown, two PiP clips are correctly detected

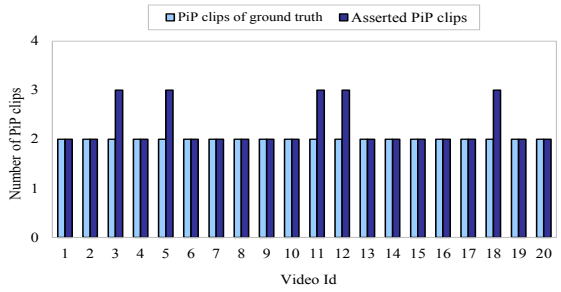

Fig.7. Comparison on the number of asserted PiP clips and the number of ground truth.

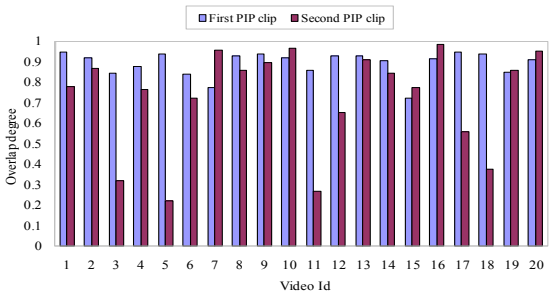

Fig.8. Evaluation of temporal localization precision on individual testing video streams.

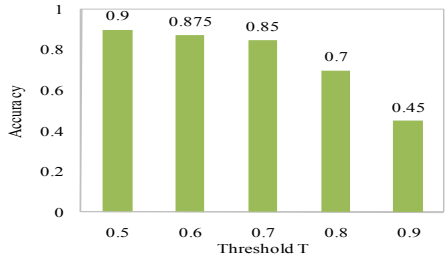

Fig.9. Evaluation on detection accuracy with different temporal localization precisions(i.e., the overlap degree)

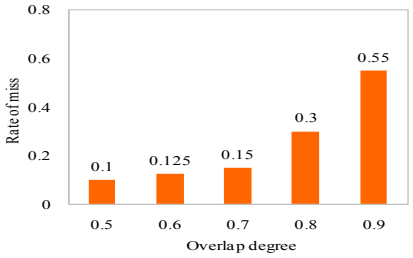

Fig.10. The evaluation on the miss rate with different temporal localization precisions(i.e., the overlap degree)

by the proposed method in most of video streams, which is consistent with the groundtruth. For four video streams, only one more PiP clip is detected, which means that only one desired PiP clip is split into two sub-clips. In next section, we will detail the discussion on PiP clip splitting.

The temporal localization precision on 20 video streams that truly contain PiP clips is individually illustrated in Figure 8. We can observe that the overlap degree is very

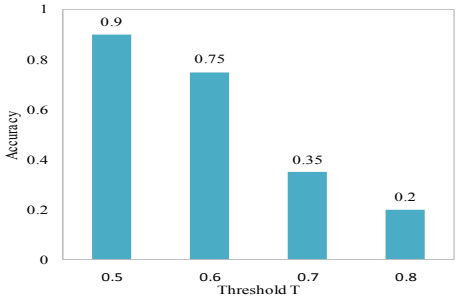

Fig.11. Evaluation on detection accuracy with different temporal localization precisions(i.e., the overlap degree)

high for most PiP clips, which means high temporal localization precision of the proposed method.

4.4. Evaluation on detection accuracy

In this section, we discuss the overall detection accuracy in detail. As discussed in last section, some desired PiP video clips are probably split to several sub-clips due to limitation of single PiP detection algorithm. Here, we first evaluate the splitting degree caused by the proposed method in an quantitative manner. From the experimental results as shown in Figure 7, we can calculate the scatter degree, which is only 0.125. This means that the proposed method can correctly localize the starting and ending boundaries of PiP clips in unbounded video streams.

For the evaluation on detection accuracy, we use the criterion (9) so as to evaluate the effect of temporal localization on detection accuracy. The experimental results are illustrated in Figure 9. As shown, while the detection accuracy is relatively low when the temporal localization is restricted to a high precision, the proposed method can achieve high detection accuracy if we relax the temporal localization precision to a relative small precision like 0.5. From the view of determining whether there are PiP clips or not, the performance is acceptable.

For the evaluation on the rate of miss, the conclusion is similar with the detection accuracy. When we relax the temporal localization precision, the rate of miss can dramatically decreases from 0.55 to 0.1 as shown in Figure 10.

For the evaluation on the false alarm, since all asserted PiP clips always have some overlap with the groundtruth, the false alarm is zero. This means that the proposed detection method can effectively distinguish the PiP clips from the background video streams.

4.5. Evaluation on single PiP clip detection

To the best of our knowledge, the detection problem of multiple PiP clips has not been reported in previous work. In order to compare with existing work, we degrade the proposed method to detecting single PiP clip by employing the criterion (9). As illustrated in Figure 11, the proposed method can still achieve high detection accuracy (i.e., 90%) if we relax the temporal localization precision to a relative small precision like 0.5, which is same to the detection accuracy of existing method [6] on the same dataset. This means that the proposed method still works even if it is degraded to detect single PiP clip.

5. CONCLUSION

In this paper, we proposed a spatio-temporal PiP detection approach to detect multiple PiP clips in unbounded video streams, which involves both the spatial and temporal localizations. In particular, the successive-group prior of both PiP and non-PiP clips is investigated on the basis of single PiP detection accuracy. Following the successive-group prior knowledge, a spatio-temporal localizing scheme is proposed to detect multiple PiP clips. The experimental results shown the proposed scheme can effectively distinguish the PiP clips from non-PiP background video stream. Both the temporal localization precision and overall detection performance is pretty fine.

6. ACKNOWLEDGEMENTS

This work was supported in part by the 973 Program (No. 2012CB316400), PCSIRT (No.IRT201206), the National Science Foundation of China (No.61202241, No.61210006, and No.61025013), the Fundamental Research Funds for the Central Universities (No.2013JBM024), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

7. REFERENCES

[1] C.-Y. Chiu, C.-S. Chen and L.-F. Chien, "A framework for Handling Spatiotemporal Variations in Video Copy Detection," *IEEE Transactions on Circuits Systems for Video Technology,* vol. 18, no. 3, pp. 412–417, 2008.

[2] X.-S. Hua, X. Chen, and H.-J. Zhang, "Robust Video Signature Based on Ordinal Measure," in *Proc. IEEE International Conference on Image Processing,* vol. 1, pp. 685 - 688, 2004.

[3] C. Kim and B. Vasudev, "Spatiotemporal Sequence Matching For Efficient Video Copy Detection," *IEEE Transactions on Circuits Systems for Video Technology,* vol.15, no. 1, pp. 127–132, 2005.

[4] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal Computer Vision,* vol. 60, no. 2, pp. 91–110, 2004.

[5] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up Robust Features," in *European Conference on Computer Vision,* vol. 3951, pp. 404-417, 2006.

[6] Z. Liu, Eric Zavesky, and N. Zhou, "AT&T Research at TRECVID 2011," in *Proc. TRECVID,* 2011.

[7] O. Bilal Orhan and J. Liu, "University of Central Florida at TRECVID 2008 Content Based Copy Detection and Surveillance Event Detection," in *Proc. TRECVID,* 2008.

[8] Z. Zhao, Y. Zhao and X. Guo, "BUPT-MCPRL at TRECVID 2011," in *Proc. TRECVID*, 2011.

[9] R. Maini and Dr. Himanshu Aggarwal, "Study and Comparison of Various Image Edge Detection Techniques," *International Journal of Image Processing, vol.3*, no.1, pp.1-11, 2009.

[10] J. Jensen, "Hough Transform for Straight Lines," *Mini-project in Image Processing,* 7th semester, 2007.