Adversarial Attacks on SIFT Based Video Summarization techniques

Vireshwar Kumar
Computer Science
Indian Institute of Technology, Delhi
viresh@cse.iitd.ac.in

Prateek Singh
Maths and Computing
Indian Institute of Technology, Delhi
mt6180787@iitd.ac.in

Shreyansh Choudhary
Maths and Computing
Indian Institute of Technology, Delhi
mt6180794@iitd.ac.in

***Abstract*—Recently our understanding of various concepts of computer vision gave rise to a wide variety of applications, be it face detection, movement detection and what not. But with the alarming threat of various adversarial attacks on such models, not only in computer vision but AI in general, a lot of research has been done in order to find out their weaknesses or try to make them more robust to previously known attacks. Below we have tried to find out how robust are state-of-the-art video summarizers vs various SIFT based adversarial attacks and find out how summaries are affected.**

***Index Terms*—SIFT, Video Summarization, Adversarial Machine Learning**

I. Introduction

Video Summarization has been a topic of interest in computer vision ever since the success of various text summarizers. The idea of a video summarizer is similar, to summarise long videos into short and meaningful chunks. Given that a very large and high-quality video takes up a lot of space and a lot of bandwidth to transfer from one place to another, a viable solution is to shorten them sufficiently. Its application is very useful, especially in summarising hours of CCTV footage, where previously countless hours of footage was spent for review which is, however, inefficient and also, one can miss a lot of key details as one easily misses important information in a blink of an eye. In such cases, a shortened version of the same footage is very helpful as a good summarizer can ignore irrelevant information and only keep the relevant chunks. However, even the state-of-the-art applications are vulnerable to various adversarial attacks. In this paper, we would look into the robustness of one such state-of-the-art video summarization method. We tried various perturbations to a video before summarising it, and noted the results. Some perturbations gave bad summarised results, while one didn't affect at all.

II. Related Work

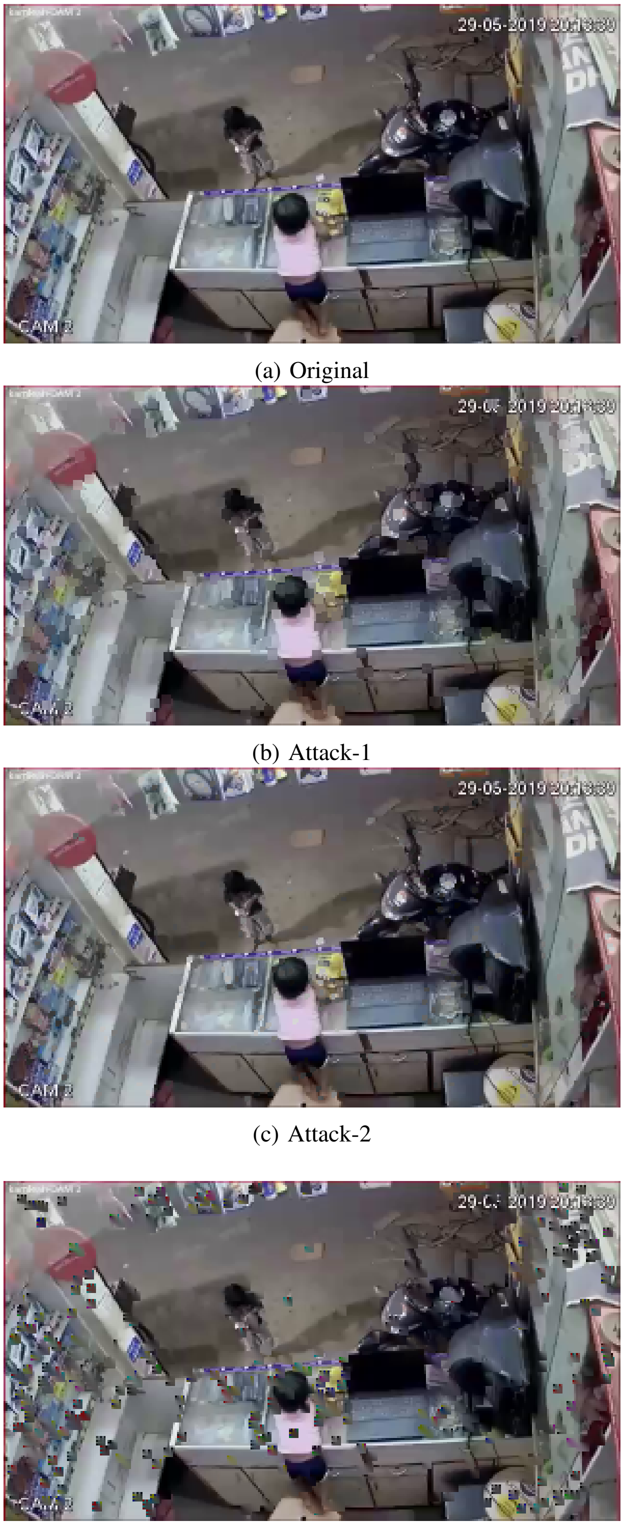
The most difficult challenge of video summarization is determining and separating the important content from the unimportant content. Most summarizers detect some important points(keypoints) from each frame, analysing their movements and then marking them important if they move significantly. As in Video Summarization technique proposed by Shruti Jadon and Mahmood Jasim (Jadon & Jasim, 2020) in their work. They use SIFT (Morel & Yu, 2011) and Image Histograms (Yu Wang, Qian Chen, & Baeomin Zhang, 1999) to detect local features and then analyse their movements in consecutive frames. If the change in pixel values is considerable, then that frame is considered important. For all the important frames, the summarizer then skims (Yihong Gong & Xin Liu, 2000) some frames in their neighbourhood so that the final video is continuous.
Another approach, which is helpful for surveillance systems (Po Kong Lai, Décombas, Moutet, & Laganière, 2016), separates foreground using background subtraction (KaewTraKulPong & Bowden, 2002), dense optical flow (Farnebäck, 2003), and gravitational clustering (Gomez, Dasgupta, & Nasraoui, 2003), associates detected foreground objects in consecutive frames through sliding temporal windows and then selects the keyframes. Other interesting work on summarization have been done in (Fu, Tai, & Chen, 2019; H. Wei et al., 2018; Ma, Lu, Zhang, & Li, 2002; Lee, Ghosh, & Grauman, 2012; Zhang, Chao, Sha, & Grauman, 2016).
Various attacks have been studied on such summarizations before, some (X. Wei, Zhu, Yuan, & Su, 2019) are direct attacks, while some (Yan, Wei, & Li, 2020) uses machine learning like reinforcement learning to do the same.

III. Proposed Attacks

We first tried to adjust the parameters of the video summarizer to work the best for CCTV footages. For this we asked people to find a summary for the videos, we have tested on and considered them as ground truth summary. We then tried to fine tune the parameters to get the best possible results from the summarizer. In this report, we will be using a white box attack to target the state-of-the-art video summarizer. We have adopted attacks on SIFT algorithm suggested in the paper by Zohaib et. al (Ali, Anjum, & Hussain, 2019). The attacks fool the summarizer by perturbing the videos such that the Matcher provides incorrect matching of the keypoints. The following are the attacks mentioned in the work by Zohaib et. al. (Ali et al., 2019) on SIFT which worked fairly well.

A. Attack-1: Blur

The first perturbation that we tried out was to blur the images at the corresponding detected key points. The idea was that if the SIFT detector would find it difficult to find



(a) Original

(b) Attack-1

(c) Attack-2

(d) Attack-3

Fig. 1: Perturbations Caused by the Attacks

the respective key points, then the corresponding matching would not be good. And as a consequence it would give us unsatisfactory results. The big problem with this attack was, athough it did manage to affect the summary, the perturbations were visible to the naked eye. And hence this is not a great adversarial attack.

B. *Attack-2: Block-2-Block*

The second attack was as Zohaib et. al. (Ali et al., 2019) named, Block-2-Block perturbation. This attack seemed much more successful as it disrupts the summary very well. In this, around every SIFT feature detected a patch (usually 3x3) is taken and blocks are then perturbed according to a threshold. This threshold is a parameter that can be fine tuned to ones liking. Unlike the other two, this perturbation seemed impossible to detect from the naked eye.

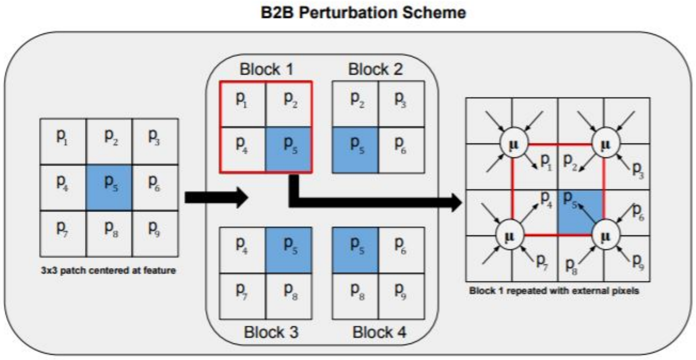

B2B Perturbation Scheme
3x3 patch centered at feature
Block 1
Block 2
Block 3
Block 4
Block 1 repeated with external pixels
μ

Fig. 2: Attack-2

C. *Attack-3: Pixel-2-Pixel*

This was the last attack we tried, and this was not very successful. In this method, we would take the average of the left and upper pixel of the detected keypoint and replace it with the current pixel. This attack was not very effective, even as an adversary. The modified pixels were too obvious to the naked eye.

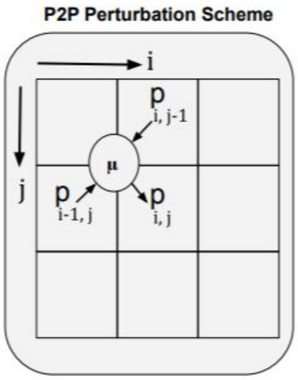


Fig. 3: Attack-3

IV. Results

The videos we used for testing were downloaded from Insecam (*Insecam, Dataset for CCTV videos*, n.d.) and trimmed to manageable sizes. We test the perturbed videos in 2 different ways. Firstly, it is important that the perturbations should not affect the video that can be easily perceived by human eye and secondly, the summarizer should fail to detect the critical frames from the required/ ground truth summary. For ground truth summary, we have asked people to provide the same for 10 small CCTV footages. Below we mention our results.

A. Degree Of Perturbation

Here we have chosen SSIM(structural similarity) (Zhou Wang, Bovik, Sheikh, & Simoncelli, 2004) as a metric for the similarity between the perturbed video frames and original frames. To present the results here, we have averaged them for all the videos.

Average SSIM scores	
Attack Type	SSIM Score
Attack-1	0.92
Attack-2	0.966
Attack-3	0.859

Here we can see that Attack-2(Block 2 Block) performs the best and the perturbations are not visible to human eye.

B. Summarizer Performances

For this work, we have used SIFT based summarizer proposed by Shruti Jadon and Mahmood Jasim . We mark the frames in the ground truth summary obtained from different people as critical and check how many of those are included by the summary of the original video and attack summary.

Summarizer performance on differently perturbed videos				
Attack Type	Original Video	Attack-1	Attack-2	Attack-3
1	0.8067	0.6992	0.7593	0.7222
2	0.99	0.972	0.962	0.921
3	0.941	0.90	0.856	0.901
4	0.8584	0.8159	0.8278	0.8398
5	0.9130	0.704	0.6393	0.5821
6	0.8144	0.7357	0.7663	0.7542
7	0.8983	0.8569	0.7974	0.8554

Further Work

There is a lot of scope for developing security for these SIFT based attacks. SIFT is a wildly used and robust technique, though it takes too much time per frame of the video. Hence, much work should be concentrated on the working time of both the summarizer and attacks. Some other attacks on this summarizer is using pepper noise on alternating frames to fool the motion threshold.

References

Ali, Z., Anjum, M. L., & Hussain, W. (2019). Adversarial examples for handcrafted features. In *Bmvc.*

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on image analysis* (pp. 363–370).

Fu, T., Tai, S., & Chen, H. (2019). Attentive and adversarial learning for video summarization. In *2019 ieee winter conference on applications of computer vision (wacv)* (p. 1579-1587). doi: 10.1109/WACV.2019.00173

Gomez, J., Dasgupta, D., & Nasraoui, O. (2003). A new gravitational clustering algorithm. In *Proceedings of the 2003 siam international conference on data mining* (pp. 83–94).

Insecam, dataset for cctv videos. (n.d.). Retrieved from `https://www.insecam.org/en/`

Jadon, S., & Jasim, M. (2020). *Video summarization using keyframe extraction and video skimming* (Tech. Rep.). EasyChair.

KaewTraKulPong, P., & Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems* (pp. 135–144). Springer.

Lee, Y. J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *2012 ieee conference on computer vision and pattern recognition* (pp. 1346–1353).

Ma, Y.-F., Lu, L., Zhang, H.-J., & Li, M. (2002). A user attention model for video summarization. In *Proceedings of the tenth acm international conference on multimedia* (pp. 533–542).

Morel, J.-M., & Yu, G. (2011). Is sift scale invariant? *Inverse Problems and Imaging*, *5*(1), 115–136.

Po Kong Lai, Décombas, M., Moutet, K., & Laganière, R. (2016). Video summarization of surveillance cameras. In *2016 13th ieee international conference on advanced video and signal based surveillance (avss)* (p. 286-294). doi: 10.1109/AVSS.2016.7738018

Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., & Yao, C. (2018). Video summarization via semantic attended networks. In *Aaai.*

Wei, X., Zhu, J., Yuan, S., & Su, H. (2019). Sparse adversarial perturbations for videos. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 8973–8980).

Yan, H., Wei, X., & Li, B. (2020). Sparse black-box video attack with reinforcement learning. *arXiv preprint arXiv:2001.03754.*

Yihong Gong, & Xin Liu. (2000). Video summarization using singular value decomposition. In *Proceedings ieee conference on computer vision and pattern recognition. cvpr 2000 (cat. no.pr00662)* (Vol. 2, p. 174-180 vol.2). doi: 10.1109/CVPR.2000.854772

Yu Wang, Qian Chen, & Baeomin Zhang. (1999). Image enhancement based on equal area dualistic sub-image histogram equalization method. *IEEE Transactions on Consumer Electronics*, *45*(1), 68-75. doi: 10.1109/30.754419

Zhang, K., Chao, W.-L., Sha, F., & Grauman, K. (2016). Video summarization with long short-term memory. *ArXiv*, *abs/1605.08110.*

Zhou Wang, Bovik, A. C., Sheikh, H. R., & Simoncelli,

E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600-612. doi: 10.1109/TIP.2003.819861