Practical Uses of A Semi-automatic Video Object Extraction System

Surya Sumpeno, Mochamad Hariadi [1] Takafumi Aoki[2]
1) Department of Electrical Engineering, Faculty of Industrial Technology ITS Surabaya Indonesia 60111, email: surya@ee.its.ac.id
2) Graduate School of Information Sciences (GSIS) Tohoku University, Japan

Abstract *- Object-based technology is important for computer vision applications including gesture understanding, image recognition, augmented reality, etc. However, extracting the shape information of semantic objects from video sequences is a very difficult task, since this information is not explicitly provided within the video data. Therefore, an application for exttracting the semantic video object is indispensable and important for many advanced applications.*

An algorithm for semi-automatic video object extraction system has been developed. The performance measures of video object extraction system; including evaluation using ground truth and error metric is shown, followed by some practical uses of our video object extraction system.

The principle at the basis of semi-automatic object extraction technique is the interaction of the user during some stages of the segmentation process, whereby the semantic information is provided directly by the user. After the user provides the initial segmentation of the semantic video objects, a tracking mechanism follows its temporal transformation in the subsequent frames, thus propagating the semantic information.

Since the tracking tends to introduce boundary errors, the semantic information can be refreshed by the user at certain key frame locations in the video sequence. The tracking mechanism can also operate in forward or backward direction of the video sequence.

The performance analysis of the results is described using single and multiple key frames; Mean Error and "Last_Error", and also forward and backward extraction. To achieve best performance, results from forward and backward extraction can be merged.

Keywords*: forward and backward semi-automatic video object extraction, performance evaluation, multiple key frames.*

1.INTRODUCTION

Nowadays, the emerging video coding standard MPEG-4 enables various content-based functionalities for new types of content-based applications [1]. MPEG-4 provides standardized ways to encode video and audio objects, and the scene description, which indicates how the objects are organized in a scene.

One of the most important innovations that MPEG-4 brings is the capability of manipulating the individual objects in an image sequence (video). To fully make use of these advanced functionalities, object-based video processing is required. The main purpose of video object extraction techniques is to obtain a semantic video object. A semantic video object corresponds to a human abstraction.

Recent developments in video object extraction research lead to two types of video object extraction technique i.e., automatic extraction (e.g., [2]) and semi-automatic extraction (e.g. [3], [4]). In automatic technique, object extraction is automatically done without user intervention. Automatic extraction technique is usually based on special characteristics of the scene or on specific knowledge (i.e. a priori information) such as colors, textures and motions [5]. The inherent problem of this technique is that it is difficult to automatically extract a semantically meaningful object, since the object may have multiple colors, textures and motions.

In semi-automatic extraction technique, user is required to provide semantic information. A semi-automatic video object extraction technique based on Learning Vector Quantization (LVQ) has been developed [6], [7], [8], [9], [10], [11], [12], [13]. This technique belongs to semi-automatic approach. It requires a key frame in which the semantic object of interest is manually given by the user at the beginning of video object extraction process.

2. METHOD

The track mechanism is based on LVQ, which provides optimal class decision for distinguishing between the object of interest and the background. LVQ codebook vectors are utilized to maintain the class of each region for tracking the semantic object. Each pixel of a video frame is represented by a 5-dimensional (5-D) feature vector integrating spatial and color features. Spatial feature refers to pixel position in 2-D coordinates, while color feature is represented by YUV color space components [13],[14].

The accuracy of video object extraction is evaluated with help of ground truth [15]. The basic idea -- which is common in most types of evaluations-- is a comparison between the algorithm generated output and some ideal version of "truth" (ground truth) [16]. Evaluation therefore consists in comparing the shape

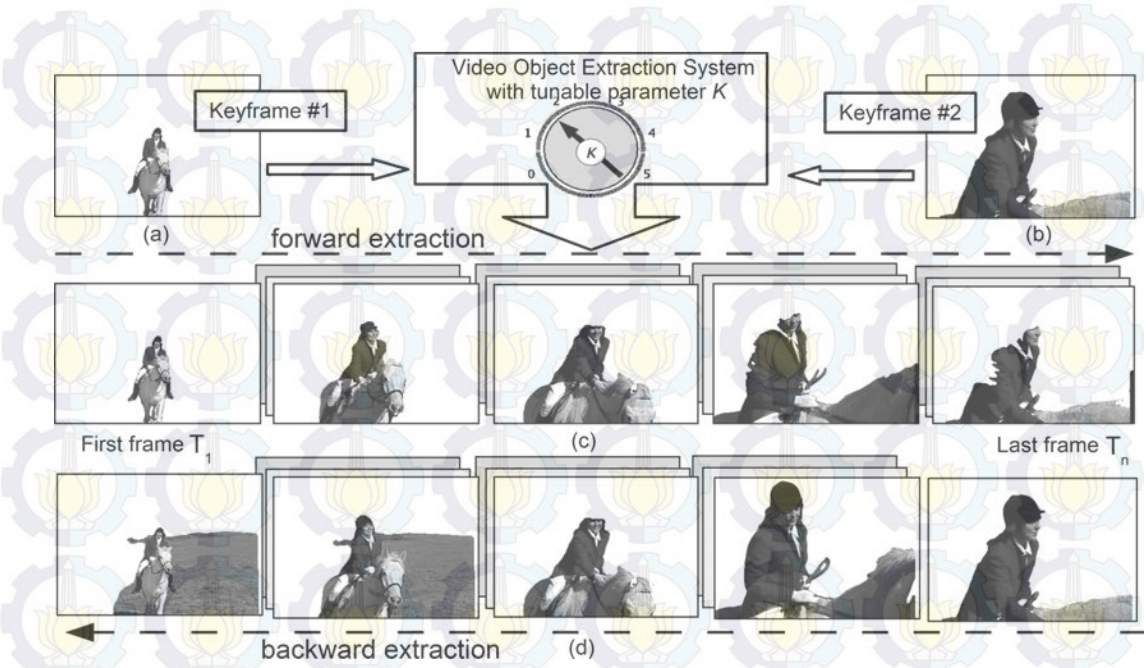

Fig. 1: Video object extraction system with tunable parameter *K* in practical uses
(a) key frame #1 (b) key frame #2 (c) forward extraction (d) backward extraction

of the reference segmentation mask (ground truth) with the shape of the ones obtained by the algorithm to evaluate. The performance of video object extraction can be viewed as a set of metrics of interest on the output of video object extraction algorithms with respect to the reference segmentation mask.

Experimental results of our video object extraction system can be evaluated using mean error, which requires all ground truth frames of full sequence to evaluate the system thoroughly. However in real uses, ground truth frames do not exist. Instead, when evaluating the error for a temporal segment, there is another measure, namely error of the last frame of temporal sequence and it is called "*Last_Error*".

Fig. 1 shows the implementation of our video object extraction system in real world. A key frame is given by user at the beginning of temporal segment (1st key frame) to start the extraction process, while a 2nd key frame (Key Frame #2) at the end of temporal segment is made by user to objectively evaluate the results of object extraction using Last_error.

Semi-automatic means user's assistance is needed at the first frame and then the system goes through the sequence forward do the extraction. When the error goes uphill above a certain threshold, the video object extraction process can be stopped, then a new key frame provided by user is inserted, and the process restarts again. This multiple key frames approach will maintain the extraction quality at a reasonable level in spite of occlusion and other changes. Fig. 2 shows expected frame by frame error using single key frame (or two key frames if 1st key frame functions as initial assistance and 2nd key frame as mean to evaluate the result using *Last_error*) and multiple key frames.

The decision that must be made when an insertion of a new key frame is necessary can be based on visual evaluation by user, who manually views the results of object extraction frame by frame.

In a lengthy sequence which consists of hundreds of frames, using our video object extraction system that has a tunable weight parameter *K* which can be set to a number of different values, hence producing different results, user will face probably thousands of possible results.

However in practical uses, user does not have to view all, instead a limited number of frames which are sampled from a full sequence of the results. By viewing samples of result, user is still able to evaluate the performance of object extraction. A simplest method is to take the last frame of results as the single sample from the full sequence of results.

Using first key frame as start point and last key frame as evaluation point, as already illustrated in Fig. 1, user has to provide two key frames to get the best possible results from our video object extraction system.

Another possible and practical approach to fully take advantage of these two key frames is forward and backward object extraction.

Instead extracting the object from frame T_1 up to frame T_n (where *n* is the number of frames) as usual using forward extraction approach, it is also possible to do backward extraction which proceeds from frame T_n down to frame T_1. In backward extraction approach, key frame at the end of temporal segment serves as initial key frame to do the object extraction process.

Yet another better result still can be obtained by merging best results from both approaches. If error of object extraction increases as process goes on leaving

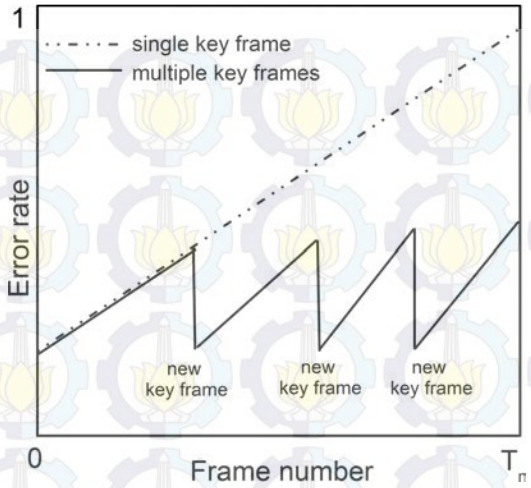

Fig. 2: Comparison of expected error rate between single key frame and multiple key frames

its start point where a key frame is provided by human, then rough figure of expected error rate from merging of both approaches can be illustrated in the Fig. 3.

It is assumed there is an in-between point T_d lies somewhere between T_1 and T_n. A merged result whose lower mean error is picked from the best result of forward extraction at T_1~ T_d and from the best result of backward extraction at T_{d+1}~ T_n .

The results of both approaches can be merged to obtain a better result i.e. lower error rate (see Fig. 3). Utilizing frame-by-frame error evaluation, the exact position of T_d can be determined, however in practical use, this evaluation cannot be done. A simplest way to merge the results is to define T_d as middle point between T_1 and T_n, i.e. $T_{n/2}$.

3. RESULTS AND DISCUSSIONS

In general, for lengthy real-world video sequences, the object of interest may undergo many changes over time. The object may become occluded, change its size, shape or color, or move too quickly, resulting in extraction errors. For practical application of this technique, it is necessary for the user to manually define the object of interest in multiple key frames to refresh the semantic information.

Tested sequences are "Foreman" and "Horse riding" which have errors in the some of their last consecutive frames. Insertion of a new 2nd key frame occurs at T=192 for "Foreman" sequence and at T=878 for "Horse Riding" sequence. To obtain best result for Foreman sequence, K_1=3.1 and K_2=2.3 are utilized, while for "Horse riding" sequence, K_1=K_2=3.6.

After inserting a new key frame and restart the video object extraction system from the new point, mean error drops as expected. Mean error of "Foreman" sequence equals to 5.90 when single key frame is used and drops to 4.58 when utilizing two key frames,

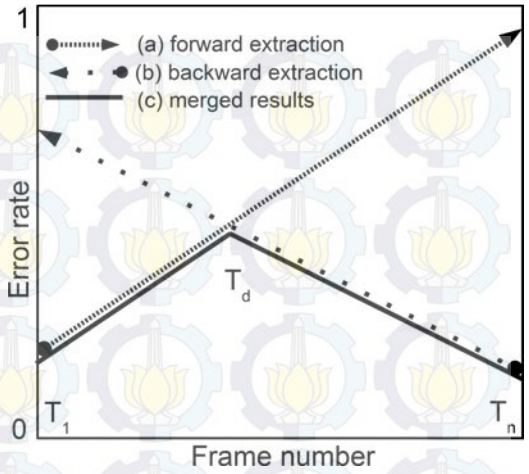

Fig. 3: Expected error rate of merged results

while mean error of "Horse riding" sequence equals to 2.55 with a single key frame, drops to 2.32 when utilizing two key frames.

For high-quality extraction, it is essential to use an appropriate parameter value of tunable parameter *K* as such that mean error will be lowest as possible. On the other hand, *Last_error* can be evaluated quickly and easily.

Experiments are done using a set of *K*'s, ranges from 1.0 to 3.6 with 0.1 as an increment value. Fig. 4 shows that *Last_error* correlates strongly with mean error, i.e. low *Last_error* always means low mean error. This trend is observed consistently for all four tested video sequences namely "Claire", "Foreman", "Horse riding and "Mother-daughter" sequence.

In a lengthy video sequence, a key frame can be taken from anywhere point and then do the forward or the backward extraction. Fig. 5 shows the frame-by-frame error rate of forward and backward extraction for "Foreman" sequence, and generally follows similar pattern --i.e. error rate tends to increase as moving forward/backward away from key frame-- as expected error rate of forward and backward extraction.

Foreman sequence is backward extracted and Fig. 6 (c) shows the best result of it. Lowest mean error of backward extraction results equals to 5.23, which is lower than the lowest mean error from the results of forward extraction which equals to 5.90. In other words, using backward extraction alone, a better result for Foreman sequence is achieved.

The mean error of merged result from forward and extraction approach equals to 4.70, which is lower than using either forward or backward extraction approach alone. The mean error of the merged result drops as expected.

The followings need to be considered in defining a key frame, because some properties of key frame have significant effect on performance of video object extraction system:

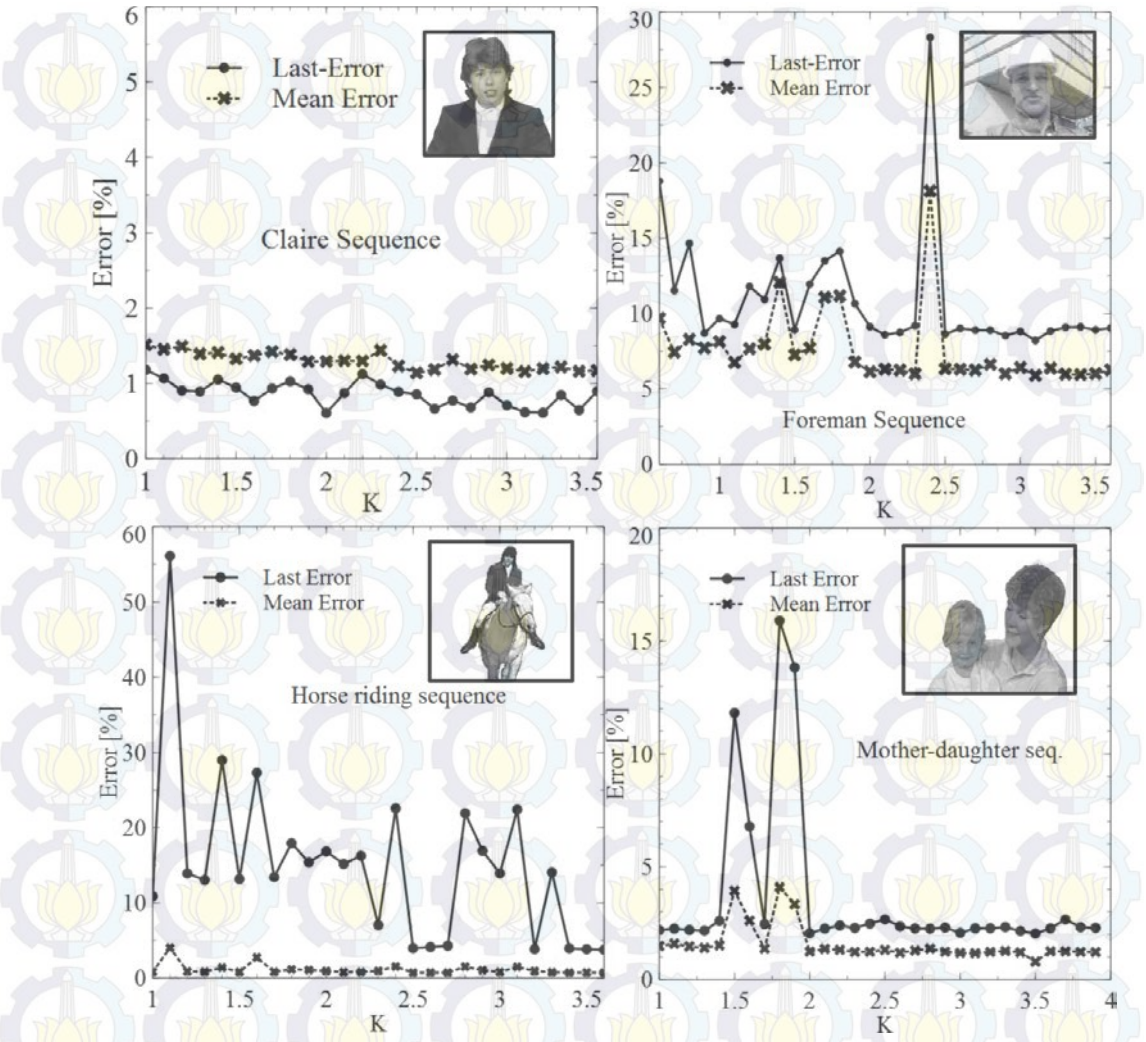

Fig. 4: Mean error and Last_Error relation

1. Definition of complete object.

Choosing a key frame whose object mask covers a complete object is recommended. For example, last frame of Horse Riding sequence which does not show the head of the horse is not a good key frame as start point to do backward extraction for full sequence and will fail at certain point. Alternatively, key frame at beginning as start point can be picked to do forward extraction, because it contains a complete object (i.e. female rider and her horse).

2. Pixel-wise accuracy.

Our video object extraction system relies on human to define semantic object. Since the video object extraction system is based on spatial and color features, it is important for user when defining a key frame to not misclassify the background as object or other way around, especially in boundary area. For example, if a key frame misclassifies part of the background as an object, while the background itself has a large region which has a color similarity with that misclassified object, then the system is likely to misclassify the object in the subsequent frames. Not only that, this error will propagate and in some cases even grow bigger in the subsequent frames.

3. Size of the object.

Sometimes video sequence can only be object extracted one-way to produce good results, either forward or backward extraction, but not both. For example, if an object is too small in the first frame, then forward extraction is not possible to produce good results. As a rule of thumb, take a frame whose size of desired object is big enough as start point (say a quarter of full screen size or nearly half would be better), then do the extraction process. In addition, when defining a key frame, for user to manually and carefully create a mask from a small object is more difficult than a bigger object.

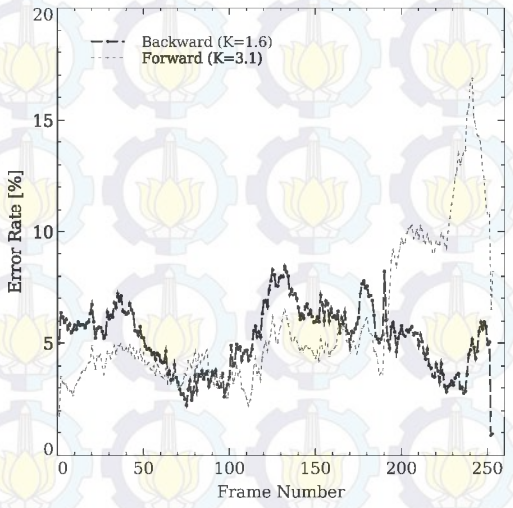

Fig. 5: Error rate of "Foreman" sequence using backward and forward extraction

4. CONCLUSIONS

In practical uses of semi-automatic video object extraction system, ground truth of all frames do not exist, therefore instead of mean error, *Last_error* can be utilized as a measure of evaluation.

Using two key frames in our semi-automatic video object extraction system, merged result with low error from best results of forward and backward extraction can be obtained. If extraction can only be done in one direction, utilizing multiple key frames can help in reducing error rate.

REFERENCES

[1] MPEG-Group. Overview of the MPEG-4 version 1 standard. ISO/IEC/JTC1/-SC29/WG11, Doc. no. 1909, Oct. 1997.

[2] Xu. H., Younis A.A., and Kabuka. M.R. Automatic moving object extraction for content-based applications. IEEE Trans. Circuits Syst. Video Technol., 14 no. 6, June 2004.

[3] S. Sun, D.R. Haynor, and Y. Kim. Semi-automatic video object segmentation using vsnakes. IEEE Trans. Circuits Syst. Video Technol., 13 no. 1, Jan. 2003.

[4] C. Toklu, A. M. Tekalp, and A. Tanju Erdem. Semi-automatic video object segmentation in the presence of occlusion. IEEE Trans. Circuits Syst. Video Technol., 10 no. 4, June 2000.

[5] A. C. Bovik. The hand book of image and video processing. Academic Press Limited, 1st edition, May 1998.

[6] Hariadi Mochamad, H. C. Loy, and Takafumi Aoki. Semi-automatic video object segmentation using LVQ with color and spatial features. IEICE Trans. Inf. Syst. Special Sect. on Recent Advances in Circuits and Systems, E88-D no. 7, July 2005.

[7] Hariadi Mochamad, Hui Chien Loy, and Takafumi Aoki. Object based video segmentation through combination of pixel position and color feature. SICE Proc. Tohoku branch 215th Workshop, Iwate, Japan, May. 2004.

[8] Hariadi Mochamad, H. C. Loy, and Takafumi Aoki. LVQ-based video object segmentation through combination of spatial and color features. IEEE Proc. TENCON 2004, Chiang Mai, Thailand, A, Nov. 2004.

[9] Hariadi Mochamad, Hui Chien Loy, and Takafumi Aoki. Integrating spatial and color features for LVQ-based video object segmentation. Proc. ITC-CSCC Matsushima, Tohoku, Japan, July 2004.

[10] M. Hariadi, A. Harada, T. Aoki, and T. Higuchi. An LVQ-based human motion segmentation. IEEE Proc. APCCAS 2002, Bali, Indonesia, II, Oct. 2002.

[11] M. Hariadi, T. Aoki, and T. Higuchi. An LVQ-based object segmentation in video sequence. IEEE Proc. of Tohoku-Section Join Convention, Yamagata, Japan, 1, Aug. 2002.

[12] M.Hariadi, A. Harada, T.Aoki, and T. Higuchi. Pixel-wise human motion segmentation using learning vector quantization. IEEE Proc. 7th International Conference on Control, Automation, and Vision, Singapore, Dec. 2002.

[13] Surya Sumpeno, Mochammad Hariadi, Takafumi Aoki. A Region-based Approach using LVQ for Semi-automatic Video Object Extraction Technique. The 8th Seminar on Intelligent Technology and Its Applications, May 9-10 2007, Surabaya Indonesia.

[14] Surya Sumpeno, Hariadi Mochamad, Koichi Ito, Takafumi Aoki. LVQ-Based Video Object Extraction and Its Performance Improvement. Journal IEIC Technical Report (Institute of Electronics, Information and Communication Engineers) Journal Code: S0532B. ISSN:0913-5685 Vol. 106; No.374 (SIS200659-70); Page.51-56(2006)

[15] Surya Sumpeno, Hariadi Mochamad and Takafumi Aoki. A Standard Performance Evaluation Tool for Video Object Extraction. The 7th Seminar on Intelligent Technology and Its Applications, May 2 2006, Surabaya Indonesia

[16] A. Cavallaro, E. D. Gelasca, and T. Ebrahimi, Objective evaluation of segmentation quality using spatio-temporal context. Proc. of IEEE ICIP, vol. 3, 2002 Sept 22-25.

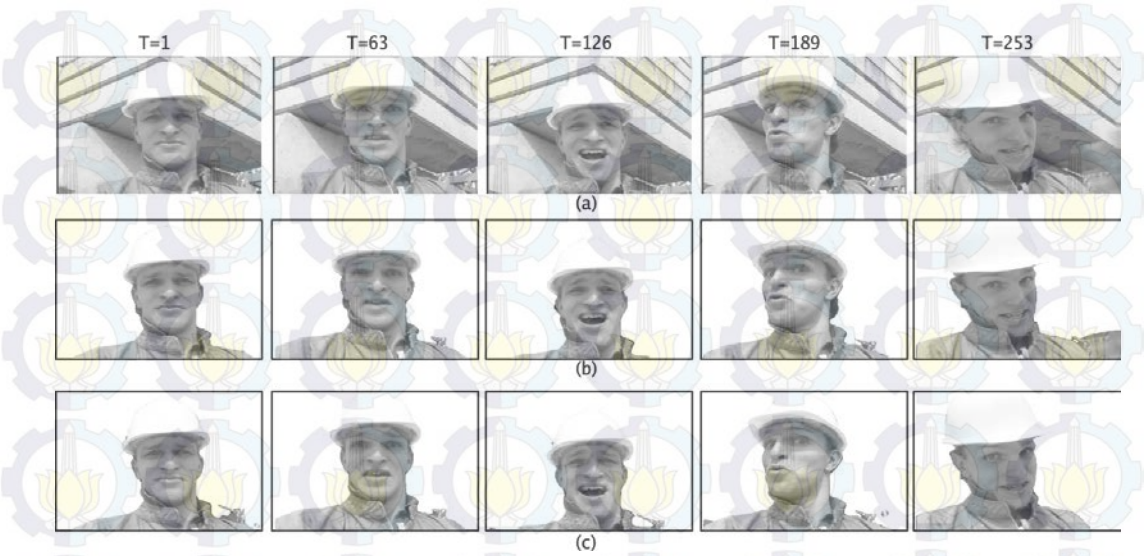

Fig. 6: Foreman sequence (a) original frames (b) best results from forward extraction (c) best results from backward extraction