INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Detection of Moving Object in Video using Space Time Context Model

Krishna Gautam[1], Surendra Choudhary [2]

[1]Govt. Engineering College, Bikaner, Rajasthan Technical University
Karni Industrial Area, Pungal Road, Bikaner, Rajasthan, India
krishnagautam3@gmail.com
[2]Assistant Prof. of Computer Science Department, Govt. Engineering College, Bikaner, Rajasthan Technical University
Karni Industrial Area, Pungal Road, Bikaner, Rajasthan, India
surendra2060@gmail.com

Abstract*: Visual tracking is a challenging problem, because the target frequently changes its appearance, arbitrarily move its location and get occluded by other objects in unhindered environments. The position changes of the target are temporally and spatially uninterrupted, in this work, a robust Spatio-Temporal structural context based Tracker (STT) is presented to complete the tracking task in unconstrained environments. The temporal context captures the historical appearance information of the target to prevent the tracker from drifting to the background in a long term tracking. The spatial context model unites contributors, which are the key-points automatically exposed around the target, to build a sustaining field. The supporting field provides much more information than appearance of the target itself so that the location of the target will be predicted more accurately. Extensive experiments on various challenging databases will demonstrate the superiority of our proposed tracker over other state-of-the-art trackers. All the simulation work will be implemented on MATLAB 2008 using image processing and generalized MATLAB toolbox as implementation platform.*

Keywords: *Spatial Context Model, Context Prior Model, Confidence Map and Weight Function,*

1. INTRODUCTION

In visual tracking, a regional context consists of a object and its proximate surrounding background inside of a decided area (See the Figure 1 regions inside the red rectangles). Hence, there exists a efficient spatio temporal relationship between the regional scenes containing the object in continuous frames. For illustration, the object in Figure 1 tolerates important occlusion which makes the object form change representatively. However, the regional context containing the object does not change much as the overall occurrence remains similar and only a insignificant part of the context region is occluded. Hence, the presence of regional context in the recent frame promotes predict the object location in the next frame. This temporally proximal information in continuous frames is the temporal context that has been currently used to object detection [1]. But, the spatial relation between an object and its regional context provides particular information about the pattern of a scene (middle column in Figure 1) which promotes discriminate the target from background when its occurrence changes much. Currently, several methods [1] employ context information to help visual tracking with demonstrated success. But, these approaches need high computational loads for feature extraction in learning and tracking phases.

The tracking problem is formed by calculating a confidence map which predicts the object location likelihood:

$$c(\mathbf{x}) = P(\mathbf{x}|o), \quad (1)$$

Where $\mathbf{x} \in R^2$ is an object location and o indicates the object display in the scene.

In the recent frame, we have the object location $\mathbf{x}^\star$ (i.e., co-ordinate of the tracked object center). The context feature class is defined as $X^c = \{\mathbf{c}(\mathbf{z}) = (I(\mathbf{z}), \mathbf{z}) | \mathbf{z} \in \Omega_c(\mathbf{x}^\star)\}$ where $I(\mathbf{z})$ shows image intensity at location $\mathbf{z}$ and $\Omega_c(\mathbf{x}^\star)$ is the neighborhood of location $\mathbf{x}^\star$. By marginalizing the joint probability $P(\mathbf{x}, \mathbf{c}(\mathbf{z})|o)$. The object location likelihood function can be calculated by:

$$c(\mathbf{x}) = P(\mathbf{x}|o)$$

$$= \sum_{\mathbf{c}(\mathbf{z}) \in X^c} P(\mathbf{x}, \mathbf{c}(\mathbf{z})|o)$$

$$= \sum_{\mathbf{c}(\mathbf{z}) \in X^c} P(\mathbf{x}|\mathbf{c}(\mathbf{z}), o) P(\mathbf{c}(\mathbf{z})|o), \quad (2)$$

Where the conditional probability $P(\mathbf{x}|\mathbf{c}(\mathbf{z}), o)$ designs the spatial relationship between the object location and its context information which promotes resolve ambiguities when the image determinations allow distinct interpretations, and $P(\mathbf{c}(\mathbf{z})|o)$ is a context prior probability which designs occurrence of the regional context. An important task in this work is to discover $P(\mathbf{x}|\mathbf{c}(\mathbf{z}), o)$ as it connetcs the gap between object location and its spatial context.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

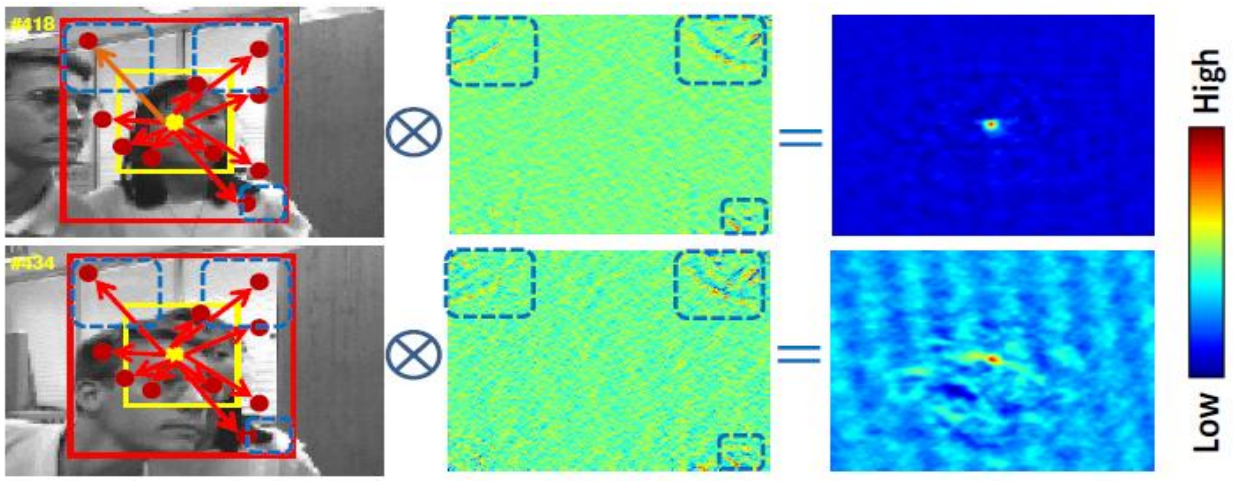


Figure 1: Left: although the target appearance changes much due to heavy occlusion, the spatial relationship between the object center (denoted by solid yellow circle) and its surrounding locations in the context region (denoted by solid red circles) is almost unchanged [1].
Middle: the learned spatio-temporal context model (the regions inside the blue rectangles have similar values which show the corresponding regions in the left frames have similar spatial relations to the target center.) [1].
Right: the learned confidence map [1].

1.1 Spatial Context Model:

The conditional probability function $P(\mathbf{x}|\mathbf{c}(\mathbf{z}), o)$ in (2) is defined as

$$P(\mathbf{x}|\mathbf{c}(\mathbf{z}), o) = h^{sc}(\mathbf{x} - \mathbf{z}), \quad (3)$$

Where $h^{sc}(\mathbf{x}-\mathbf{z})$ is a function with respect to the correspondent distance and direction between object location **x** and its regional context location **z**, thereby encoding the spatial connection between an object and its spatial context.

1.2 Context Prior Model:

In equation (2), the context prior probability is simply designed by

$$P(\mathbf{c}(\mathbf{z})|o) = I(\mathbf{z})w_\sigma(\mathbf{z} - \mathbf{x}^\star), \quad (4)$$

Where $I(\cdot)$ is image intensity that shows occurrence of context and $w_\sigma(\cdot)$ is a weighted function represented by

$$w_\sigma(\mathbf{z}) = ae^{-|\mathbf{z}|2/\sigma 2}, \quad (5)$$

Where a is a normalization constant that digest $P(\mathbf{c}(\mathbf{z})|o)$ in equation. (4) to range from 0 to 1 that satisfies the signification of probability and σ is a scale parameter.

2. RELATED WORKS

Research on face recognition from video has intensified throughout the last decade. In traditional face image acquisition settings, such as passport agencies or police stations, nuisance variables ranging from head pose to facial expression are controlled. In contrast, video surveillance systems cannot be as intrusive, so the activities of the recorded individuals and the effects of the environment can vary representatively [2, 3]. Numerous performance evaluation efforts have demonstrated that face recognition algorithms that operate well in controlled environments tend to suffer in surveillance contexts. These issues have motivated the development of face recognition algorithms that draw from the wealth of information provided by videos to compensate for the poor viewing conditions encountered in uncontrolled viewing scenarios. Specifically, assert that videos afford three useful properties that [10]:

1. A set of estimations - a video sequence contains multiple images of the same face that can potentially show how it appears under different conditions.
2. Temporal dynamics - videos contain temporal information that motionless images do not possess.
3. 3D information - in an extension to the first property, an order of video frames can show the same object from a number of different angles, i.e. 2D videos implicitly contain 3D geometric information. As well, neurological information proposes that humans exploit these properties by using both the structure of facial features and idiosyncratic facial changes to recognize others.
4. Temporal dynamics perform an especially strong role in the recognition of familiar people.

On the other hand, the following nuisance elements can mount in unconstrained face recognition applications [10]:

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

i. Pose variation - uncontrolled cameras can register non-ideal face shots from a variety of angles, causing the similarity between pixel locations and points on the face to be unlike from image to image.
ii. Illumination variation - a particular may pass underneath lights with a range of relative location and intensities throughout the course of one or more videos, so that the surface of the face occurs different at different times.
iii. Expression variation - the occurrence of the face changes as the facial expression varies.
iv. Scale variation - the face will obtain larger or smaller regions in the video frames as it proceeds towards or away from the camera, and, in the worst case, the spatial clarity of the face can decrease to the point where it get unrecognizable.
Spatial clarity can also depend on the properties of the camera, like the depth of field of its lens.
v. Motion blur - significant blur can grey the face if the camera orientation time is set too long or the head moves rapidly.
vi. Occlusion - objects in the environment can block parts of the face, making the tasks of recognizing the face and representative it from the background more difficult.
These elements may cause the differences in occurrence between distinct shots of the same person to be greater than those between two people viewed under similar situation. But pose and illumination are traditionally viewed as two of the most challenging nuisance elements, some of the other elements listed above have nearly as significant of an impact on face identification performance in uncontrolled contexts. These properties and problems are well known through the speculative, commercial and governmental sectors. The techniques are broadly classified into two groups depending on which video properties they utilize, as shown in Figure 2. First is *Set-based approaches* [2-9] regard videos as unordered group of images and take advantage of the multitude of estimations, where as *sequence-based approaches* explicitly apply temporal information to enlarge efficiency or enable identification in poor viewing conditions. Although set-based approaches do not rely on the ordering of face images, they employ the quantity and variety of recognizing to achieve robustness to degraded viewing situations. These procedures differ in terms of whether they fuse information over the recognizing before or after matching. Prior to matching, knowledge can be fused across images at the data or feature levels. Super-resolution methods operate at these levels to increase the resolution of the face.

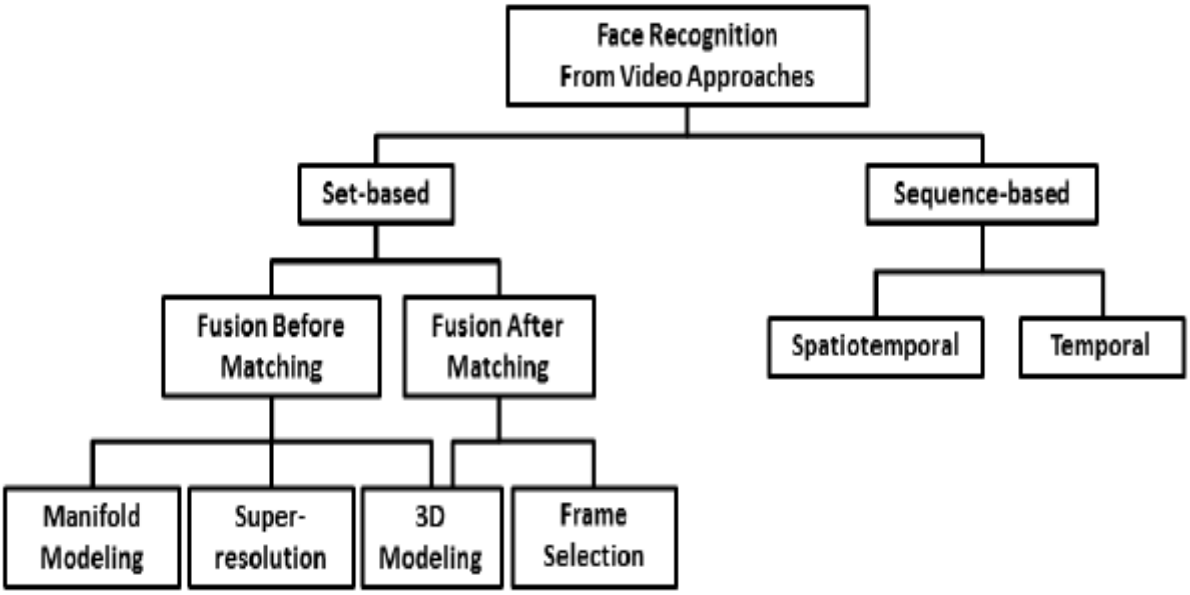


Figure 2: Taxonomy of the Face Recognition from Video Literature [10]

3. PROPOSED METHODOLOGY

1. Addition of location path of frames to MATLAB
2. Calling and inputting of frames into MATLAB.
3. Initialization of parameters for e.g. Rectangle size, extra area surrounding the target, the learning parameter (rho), initial scale ratio, Lambda, number of average frames and alpha.

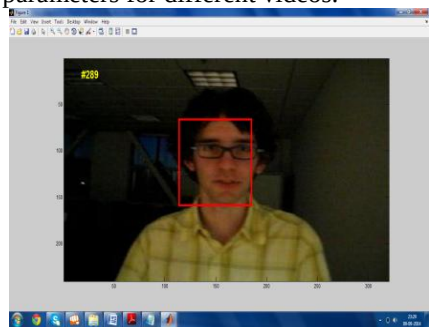
INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

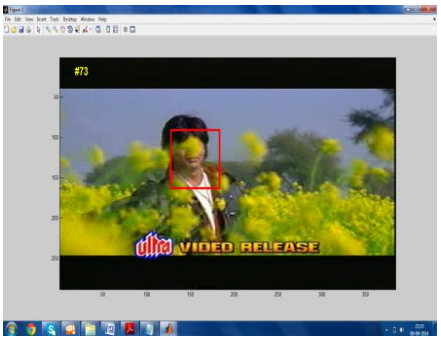
4. Calculation of context region, centre of the target, initial size of the target.
5. Computation of confidence map which estimates the object location likelihood.
   $C(\mathbf{x}) = P(\mathbf{x}|o)$
   Where, **x** is an object location and o denotes the object present in the scene.
6. Transform of confidence map into frequency domain.
7. Computation of weighted window function.
8. Normalization of weighted window function.
9. Initialization of a loop in accordance to the number of frames at the added location.
10. Calculation of sigma in accordance with initial size of the target.
11. Updating of scale in accordance with sigma.
12. Updating of weighed window function in accordance with updated scale.
13. Normalization of updated weighted window function.
14. Loading, reading and getting of 3-D matrix of 1st frame.
15. Conversion of 3-D matrix into 2-D matrix of 1st frame.
16. Computation of spatial context prior model in accordance with centre f target, size of context region and updated weighted window function with 1st frame.
17. Updating and fast learning of spatial context model.
18. Conversion of spatial context model into spatial temporal context model, for 1st frame.
19. Updating the initial size of the target according to updated scale factor.
20. Updating of centre of the target according to updated size of the target.
21. Visualization of each frame with targeted object (by red colour box).
22. From rest of the frames, repetition of steps from 4 to 15.
23. Calculation of response of the confidence map at all locations.
24. Finding of maximum response of confidence map so as to find the current location of object.
25. Updating of scale according to maximum response of confidence map.
26. Repetition of steps from 16 to 20.

4. EXPERIMENTAL RESULTS

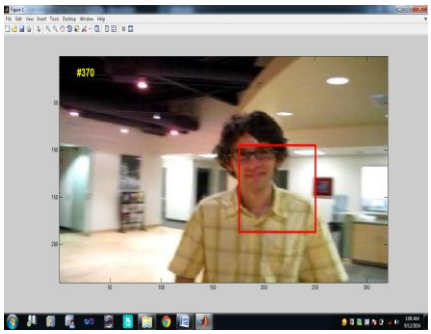
A data base of 18 videos has been taken for proposed algorithm. All the implementation work is done on MATLAB R2008a using image processing and generalized tool box. Two output parameters i.e. success rate and CLE (centre location error) has been taken for evaluation the performance of proposed algorithm. The centre location error (CLE) and success rate (SR), both computed based on the manually labeled ground truth results of each frame. The score of success rate is defined as score = area(R_t ∩ R_g)/area(R_t ∪ R_g) , where R_t is a tracked bounding box and Rg is the ground truth bounding box, and the result of one frame is considered as a success if score > 0.5. The CLE is computed as the distance between the predicted centre position and the ground truth centre position. Tracker Position is shown by a red colour rectangle. As target object moves, it moves accordingly. Our tracker computes 94.99% average success rate against other tracking algorithm. Screen shot of different videos with tracker are shown below. Also, a table is given which presents both output parameters for different videos.



(a)

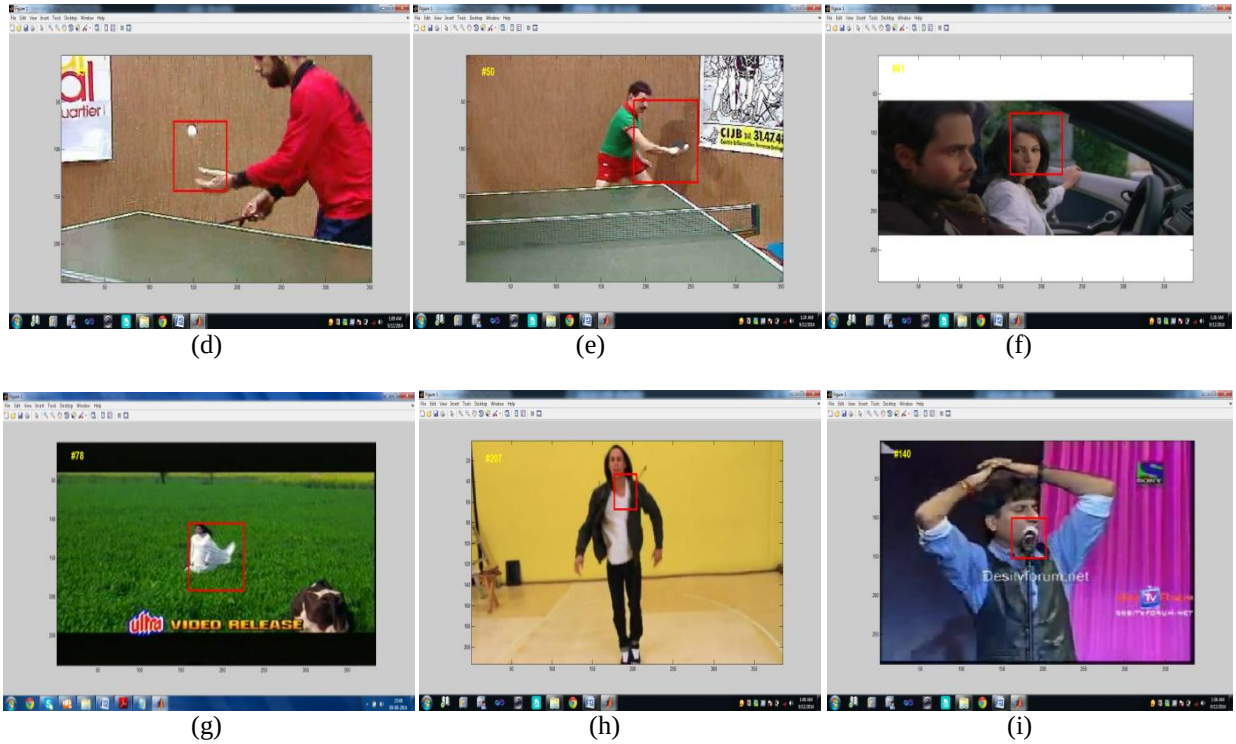


(b)



(c)

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....



(d) (e) (f)

(g) (h) (i)

Figure 3: Screenshots of our tracking approach

Table 1 Value of Success rate and CLE as output parameter

Video Sequences	Success Rate in %	Centre Location Error (CLE) in Pixels
data	97.29	0
data1	100	0
data2	98.44	0
data3	100	0
data4	100	0
data5	97.96	0
data6	74.89	0
data7	89.56	0
data8	99.20	0
data9	81.81	0
data10	100	0
data11	97.77	0
data12	95.26	0
data13	84.64	0
data14	100	0
data15	99.62	0
data16	95.31	0
data17	98.23	0
Total	1709.98	0

$$Average\ Success\ Rate\ = Total\ Success\ Rate \div Total\ Number\ of\ Video\ Sequences \quad (2)$$

$$Average\ Success\ Rate = 1709.98 \div 18$$

$$Average\ success\ Rate = 94.99\% \quad (3)$$

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Table 2: Comparisons between success rates and CLC of our proposed method and other techniques

Tracking Techniques	Average Success Rate	Average CLR
SMS	35	79
Frag	35	63
SSB	45	54
LOT	35	70
IVT	49	84
OAB	52	43
MIL	52	43
VTD	49	58
LIT	40	62
TLD	62	78
DF	53	52
MTT	59	80
Strunk	75	19
ConT	62	42
MOS	26	103
CT	62	29
CST	60	54
LGT	68	22
STC previous method(2013)	94	8
STC proposed method(2014)	94.99	0

5. CONCLUSION AND FUTURE WORK

In this paper, we present a simple yet fast and robust algorithm which exploits spatio-temporal context information for visual tracking. The Proposed algorithm performs evenly better as compared to existing methods. The algorithm is tested for data base of more than 15 videos. The evidence of the better performance of the algorithm is optimized value of success rate and CLE between ground truth size of frame and tracked object size of frame. Experimental results show that our algorithm is capable of improving both performance parameter i.e. success rate and CLE. This work has described a method for space time context model estimation using space context model and regularly updated window function. In future, an algorithm can be developed which can also identify, that the moving object is a human being or not. This can be possible by combining the skin detection technique with the proposed methodology.

REFERENCES

[1] Kaihua Zhang, Lei Zhang, Ming-Hsuan Yang, and David Zhang, "Fast Tracking via Spatio-Temporal Context Learning" http://arxiv .org/abs/1311.1939v1November 2013.

[2] J.Kwon and K.M.Lee, "Visual tracking decomposition," in CVPR, pp. 1269–1276, 2010. 2, 10, 11

[3] J.Kwon and K. M. Lee, "Tracking by sampling trackers," in ICCV, pp. 1195–1202, 2011. 2

[4] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," PAMI, vol. 33, no. 11, pp. 2259–2272, 2011. 2, 10, 11

[5] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in VPR, pp. 1305–1312, 2011.

[6] L. Sevilla-Lara and E. Learned-Miller, "Distribution fields for tracking," in CVPR, pp. 1910–1917, 2012. 2, 10, 11

[7] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking," in CVPR, pp. 1940–1947, 2012. 2, 10, 11

[8] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *CVPR*, pp. 2042–2049, 2012.

[9] L. Cehovin, M. Kristan, and A. Leonardis, "Robust visual tracking using an adaptive coupled-layer visual model," *PAMI*, vol. 35, no. 4,pp. 941–953, 2013.

[10] JEREMIAH R. BARR, KEVIN W. BOWYER, PATRICK J. FLYNN, SOMA BISWAS, "FACE RECOGNITION FROM VIDEO: A REVIEW", International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing Company, 2012.