Neural Network Model for Face Recognition from Dynamic Vision Sensor*

Fedor Shvetsov [0000-0001-5112-0430], Anton Konushin [0000-0002-6152-0021], and Anna Sokolova [0000-0001-8777-2035]
{fedor.shvetsov, anton.konushin, anna.sokolova}@graphocs.cs.msu.ru

Lomonosov Moscow State University

Abstract. In this work, we consider the applicability of the face recognition algorithms to the data obtained from a dynamic vision sensor. A basic method using a neural network model comprised of reconstruction, detection, and recognition is proposed that solves this problem. Various modifications of this algorithm and their influence on the quality of the model are considered. A small test dataset recorded on a DVS sensor is collected. The relevance of using simulated data and different approaches for its creation for training a model was investigated. The portability of the algorithm trained on synthetic data to the data obtained from the sensor with the help of fine-tuning was considered. All mentioned variations are compared to one another and also compared with conventional face recognition from RGB images on different datasets. The results showed that it is possible to use DVS data to perform face recognition with quality similar to that of RGB data.

Keywords: DVS, Face Recognition, Data Simulation

1 Introduction

In recent years, a new type of camera is gaining popularity, Dynamic Vision Sensor (DVS). While in traditional cameras, information is recorded with a certain fixed frequency (usually 25–30 times per second), the dynamic vision sensor records only the fact of a change in the level of illumination in a pixel if it exceeds a certain threshold. Thus, these cameras operate according to the principle of the human eye, which responds only to changes. This approach allows to get rid of a large amount of redundant static data, focusing only on dynamic events. Such sensors have several advantages: high speed (which allows to catch very fast events in small details), low power and memory consumption (an important feature for embedded systems, where there is no way to place a large battery and hard drive), high sensitivity (a key property for recording under extreme light conditions).

* Publication is supported by RFBR grant 18-08-01484.

Such cameras are quite expensive and do not have high resolution, however, due to the rapid development of technologies in this area, it is necessary to develop algorithms for solving various applied problems, such as 3D reconstruction, detection and tracking of objects [1]. One of these tasks is person identification, since these cameras are currently often used in video surveillance systems. There are various ways to recognize a person by frame, for example, by the walk [2] [3] [4] or by face. If the face is distinguishable in the frame and has a sufficient size, it makes sense to recognize person by it.

Nowadays, systems that perform face recognition are very relevant, since they implement the most effective way of contact-less identification of a person. They are used in security systems, bank card verification, people mark-up, forensics, online payments etc. Face recognition problem can be decomposed into several sub-tasks: finding the face in the image, normalizing the found face and, finally, identifying the person. In this paper we propose a new method for face recognition using data from DVS sensor.

2 Related Work

2.1 Detection

The problem of detecting faces is one of the special cases of the detection problems, but it has its own specificities. The human face has distinctive features, which were searched and analyzed in first approaches in this area. A big breakthrough was [5], which used haar filters to find faces using cascades of detectors. However, such algorithms did not provide stability, since faces had great variability due to different lighting and viewing angles.

Then the partially-deformed models [6] were proposed, which aimed at solving detection problem. However, these methods were computationally costly and required complex markup for training.

Similarly to most of computer vision tasks, the detection problem can be solved by the deep learning methods, the popularity of which has grown significantly after the work [7]. These methods have also been successfully applied to the task of detecting faces, for example, in [8]. Being also based on this approach, the work MTCNN [9] uses three light neural networks to find faces in the image.

2.2 Recognition

The face recognition problem has been of interest to the scientific world for a long time. The first systems for solving this problem were developed back in 1964 [10]. Since then, the level of quality of this technology has greatly increased, and modern algorithms are able to distinguish people's faces better than the people themselves [11].

Various methods were used to solve the face recognition problem, and these methods have changed greatly over time. The first algorithms attempted to distinguish between faces by finding distinctive features such as eye color, face proportions, etc. [12]. The work [13] has made a great contribution to the development of methods by using the similarity of eigenvectors for faces (eigenface). However, in general, the majority

of modern methods [14] try to recognize faces by creating embeddings for them. These methods have become especially popular after the widespread use of convolutional neural networks [15]. The same approach is used in the popular work [16].

2.3 Reconstruction

One of the key aspects of this work is an algorithm for reconstructing frames from the stream of events of a dynamic vision sensor. The algorithm was proposed by [17]. It also apply artificial neural networks. In this case, recurrent neural networks are used, the main feature of which is the ability to memorize the state obtained by processing the next element of the sequence and use it for further calculations. In this algorithm, the neural network receives at the input of stream of events from the dynamic vision sensor for a certain period of time, and the model reconstructs an image that visually looks like grayscale image.

3 Proposed method

3.1 Formal problem

The face recognition problem can be considered in two equivalent forms: identification and verification. In this paper verification form was chosen.

Data from the dynamic vision sensor comes in the form of set of events.

Event — $(x,\ y,\ ts,\ p)$, where $x,\ y \in \mathbb{Z}, x \in [0, N], y \in [0, M]$ are the coordinates of the pixel in the matrix $N \times M$, $ts \in \mathbb{R}$ — timestamp, $p \in \{-1, 1\}$ — the polarity of the change (the brightness in the pixel decreased/increased by a given threshold).

Algorithm Input: sets of events T_1 and T_2 received from the dynamic vision sensor.

Algorithm Output: $A \in \{0, 1\}$: $A = 1$ if the sets T_1 and T_2 describe one person, $A = 0$ if different.

3.2 Our method

Since the main source of information in computer vision tasks is usually an image or a sequence of images, but not the event stream obtained by DVS, it is necessary to convert the stream of events into visual representation. Such a visualization can be made in different ways. The most simple one is setting time marks and counting events occurred between and visualizing it in gray-scale. However, it turns out, that this approach does not provide with satisfying quality and detectors could not find faces on such images. Thus, the reconstructions with neural network [17] were used for visualizations which yields much better results (see Fig. 1).

The proposed basic method works as follows (see Fig. 2): first, the stream of events from the sensor is reconstructed into frames with model [17], then the faces are located in frames using the detector [9], then the neural network calculates the internal representations for them in the form of vectors $\mathbf{f_1}, \mathbf{f_2} \in \mathbb{R}^n$ [16], then the proximity of these vectors is determined using the cosine distance (1). If the proximity is higher than the specified threshold, then 1 is predicted, otherwise – 0.

Fig. 1. Simple visualization (left), neural net reconstruction (right)

$$\cos(\mathbf{f_1}, \mathbf{f_2}) = \frac{\mathbf{f_1}\mathbf{f_2}}{\|\mathbf{f_1}\|\|\mathbf{f_2}\|} = \frac{\sum_{i=1}^{n} \mathbf{f_1}_i \mathbf{f_2}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{f_1}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{f_2}_i)^2}} \tag{1}$$

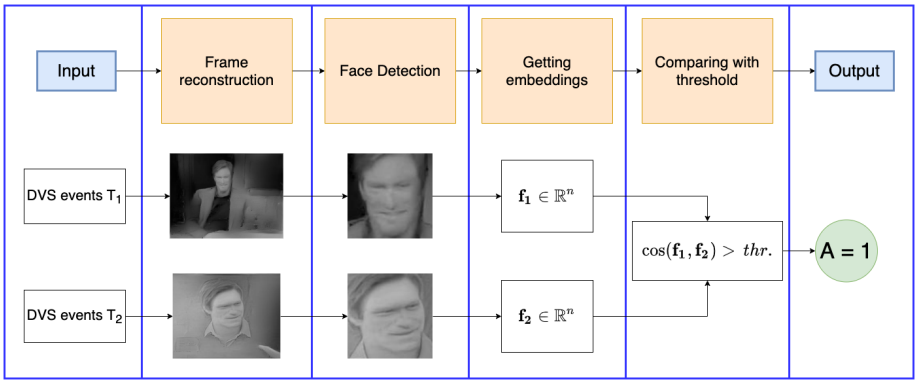

Fig. 2. Overall pipeline of method

4 Data simulation

The dynamic vision sensor is a fairly new type of cameras, and very few datasets have been recorded for it so far. As far as we know there are no publicly available datasets for a face recognition task. Therefore, it was proposed to use the collections of color videos collected for the face recognition task and simulate dynamic vision sensor data from them. It can be done in two steps: firstly, the intermediate values in each pixel are interpolated between two neighboring frames and secondly, at each point the change in

intensity between adjacent interpolated frames is compared, and if this change exceeds the threshold, an event is generated. As you can see in fig. 3, the results of real and simulated event streams are very similar. This gives us the opportunity to assume that studies conducted on simulated data will be fairly well transferred to real data.

Fig. 3. Data from sensor (left) simulated frame (right)

Since the linear interpolation of intermediate frames is not very fair leading to blurred frames, it was proposed to improve the simulation method by using better approximation. To do this, it was decided to use the results of [18], which creates a slow motion effect, and incorporate them into the reconstruction process. This approach uses the creation of intermediate frames to simulate dynamic vision sensor data from color video sequences, thereby smoothing visualizations. A variation with the creation of one intermediate frame was used. In fig. 4 the visual difference in the images presented.

Fig. 4. Simple neural net reconstruction (left) and with intermediate frame (right)

5 Experimental Evaluation

5.1 Datasets

YouTubeFaces. The main dataset that met the criteria for simulation was YouTube-Faces Dataset [19], which consists of videos collected from the YouTube, each of which contains a specific person. Its main advantage is a large number of people. The collection consists of 3425 video sequences containing information about 1595 people. Due to the large number of subjects in this dataset, it was possible to apply the neural network fine-tuning. Two-thirds of the collection was held-out for the training set, and one-third for testing, where all videos with a specific person were fully included in either the first or second group. Fine-tuning on the training set was performed where original network was trained on VGGFace2 [20] collection.

ChokePoint. In addition, it is proposed to use dataset obtained under conditions similar to real scenarios of using a dynamic vision sensor as a test set. For this, the ChokePoint [21] dataset was selected. In this dataset, 48 video sequences were recorded containing 40 people passing through the entrance to the room. Along with the video, frame-by-frame mark-up of a person in the frame is provided. The viewing angles of the cameras used to record the dataset are similar to the same angles in video surveillance systems, which reflects the possible location of the dynamic vision sensor designed to solve this problem.

GML DVS. In order to check the portability of the created model for real data obtained from a dynamic vision sensor, a small dataset of eight video sequences containing eight people was captured. 80 faces were automatically found by face detector and manually labelled.

5.2 Results

To evaluate the quality of the proposed method, we make a verification experiment selecting the pairs of objects and comparing their similarity with some threshold to decide if they belong to the same person or not. Setting different thresholds to distinguish faces we can obtain AUC metric which is the area below ROC curve and that can be a great indicator of general performance of the model. The method is tested against verification on RGB images when possible. Variations with advanced reconstruction and uses of fine-tuning on those reconstructions are examined. The results are presented in Table 1 and 2.

We can see that recognition results on DVS reconstructions are quite similar to those on RGB images and that fine-tuning enables us to improve quality of a model. Furthermore, this fine-tuning allows to enhance performance on data obtain from the real DVS sensor which was also quite good comparing to simulated data proving the portability of model.

Table 1. AUC metric for simulated images

Experiment	YTF	ChokePoint
RGB images	0.958	0.996
Reconstructions	0.922	0.948
Reconstructions + fine-tuning	**0.928**	0.961
Advanced Reconstructions	0.921	0.952
Advanced Reconstructions + fine-tuning	0.922	**0.962**

Table 2. AUC metric for GML DVS

Experiment	AUC
Reconstructions	0.931
Reconstructions + fine-tuned	0.936
Advanced Reconstructions + fine-tuned	**0.954**

6 Conclusion

This paper explores the possibility of constructing a solution to the face recognition problem based on dynamic vision sensor data. It implements the basic solution method using a neural network model. The results show that we can apply existing methods to solve this task at a level similar to that of the RGB images. Uses of simulated frames provided a great way to improve performance of the model which is very helpful due to the scarce amount of real data.

References

1. C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. *Proceedings of the IEEE*, 102(10):1470–1484, 2014.
2. Anna Sokolova and Anton Konushin. Human identification by gait from event-based camera. In *2019 16th International Conference on Machine Vision Applications (MVA)*, IEEE Xplore Digital Library, pages 1–6. IEEE, 2019.
3. Anna Sokolova and Anton Konushin. Pose-based deep gait recognition. *IET Biometrics*, 8(2):134–143, 2018.
4. Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
5. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.
6. P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2241–2248, 2010.
7. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

8. Haoxiang Li, Zhe Lin, Xiaohui Shen, and Jonathan Brandt. A convolutional neural network cascade for face detection. pages 5325–5334, 06 2015.
9. Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.
10. Jan Bergstra Karl de Leeuw. The history of information security: A comprehensive handbook. page 264–265, 2007.
11. P. Jonathon Phillips and Alice J. O'Toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1):74 – 85, 2014.
12. Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. *IEEE transactions on pattern analysis and machine intelligence*, 15(10):1042–1052, 1993.
13. Matthew Turk and Alex Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*, pages 586–587, 1991.
14. Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.
15. Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.
16. Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.
17. Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020.
18. Huaizu Jiang, Deqing Sun, Varan Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.
19. L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, 2011.
20. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.
21. Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 81–88. IEEE, June 2011.