The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop “Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine”
PSBB25 , 9–11 June 2025, Moscow, Russia

Cooperative Face Liveness Detection from Optical Flow

Artem Sokolov[1, 2], Mikhail Nikitin[2], Anton Konushin[3, 1]

[1] M. V. Lomonosov Moscow State University, Moscow, Russia
[2] Tevian, Moscow, Russia – (artem.sokolov, mikhail.nikitin)@tevian.ru
[3] AIRI, Moscow, Russia – konushin@airi.net

Keywords: Face Liveness Detection, Cooperative Liveness detection, Video Liveness, Face Anti-spoofing, Optical Flow.

Abstract

In this work, we proposed a novel cooperative video-based face liveness detection method based on a new user interaction scenario where participants are instructed to slowly move their frontal-oriented face closer to the camera. This controlled approaching-face protocol, combined with optical flow analysis, represents the core innovation of our approach. By designing a system where users follow this specific movement pattern, we enable robust extraction of facial volume information through neural optical flow estimation, significantly improving discrimination between genuine faces and various presentation attacks (including printed photos, screen displays, masks, and video replays). Our method processes both the predicted optical flows and RGB frames through a neural classifier, effectively leveraging spatial-temporal features for more reliable liveness detection compared to passive methods.

1. Introduction

Face recognition (FR) algorithms are used to solve a wide range of practical tasks that require person identification, such as face-payment and crossing point control. Being part of identification systems which are under the threat of dealing with intruders, makes FR algorithms an object for representation attacks. The problem is that FR models treat real and spoof faces, that can be depicted on a piece of paper or a screen, in the same way. Due to this feature, if the identification system is not protected from such attacks, the presence of any person can be easily simulated. To prevent identity substitution during biometric identification, the face liveness detection algorithms are used to determine whether the face in the image is real or not.

The basic face liveness detection methods work with single input image (single-shot), but they are not always able to accurately recognize attacks on the identification systems when high-quality fake faces are used. To increase the accuracy of spoofing classification, the algorithms that take video frames as an input can be utilized (multi-shot). Such an approach gives an opportunity to take into account not only static information about face texture and surrounding, but also temporal information about face movement. To further improve the accuracy, the input video can be recorded in cooperative mode, when a user is asked to follow a set of instructions while filming. Imposing such restrictions on the process of data collection may increase total time of person identification, but significantly reduces the likelihood of a successful attack on the system.

In this work we propose the novel cooperative multi-shot facial liveness detection method. The input videos for the algorithm should be taken following a specific scenario, in which a person is slowly moving frontal-oriented face closer to the camera between two predefined checkpoints (”approaching face” scenario). The optical flow detector is used as the main component of video sequence preprocessing. The novelty of the proposed method lies in the usage of a predicted facial optical flow of the input video, which is filmed by the ”approaching face” scenario, for further real/fake classification. The use of optical flow proved to be particularly effective with the chosen video scenario, as this combination gives an opportunity to extract information about face volume. Also, the predicted optical flow is passed to the classifier along with an RGB frame from the video in order to take in account both static and temporal information.

2. Related Work

The field of video-based liveness detection has a long history of research. Some of the early approaches were focused on analyzing specific facial movements, such as eye movement and blinking patterns (Jee et al., 2006) (Nikitin et al., 2019) or lips motion (Kollreider et al., 2007). While being effective against simple spoofing attempts, these methods proved to be vulnerable to more sophisticated attacks using masks with pre-cut openings in the targeted regions.

Another group of video liveness classification approaches is based on adaptation of existing single-shot methods to video scenario. For example, (Liang et al., 2024) introduced a quality assessment module, which gives an opportunity to weight predictions of single-shot model across frames, giving higher importance to predictions on frames, that are easier to classify for single-shot model. However, such methods are not able to fully exploit temporal information as they process frames independently.

The emergence of vision transformers enabled new architectures for video analysis. Works like (Ming et al., 2022) and (Marais et al., 2023) employed temporal attention mechanisms to capture inter-frame relationships. These approaches provide high classification accuracy, though at significant computational cost.

The more efficient approaches are focused on frame aggregation. One of such algorithms was proposed by (Parkin and Grinchuk, 2020). In this method authors combined optical flow detector with rank pooling (Fernando et al., 2017) to extract features from input videos for classification. Another algorithm was proposed by (Muhammad et al., 2022), in which stabilized frame averaging was used to compute the sequence aggregation for further liveness classification.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-269-2025 | © Author(s) 2025. CC BY 4.0 License.

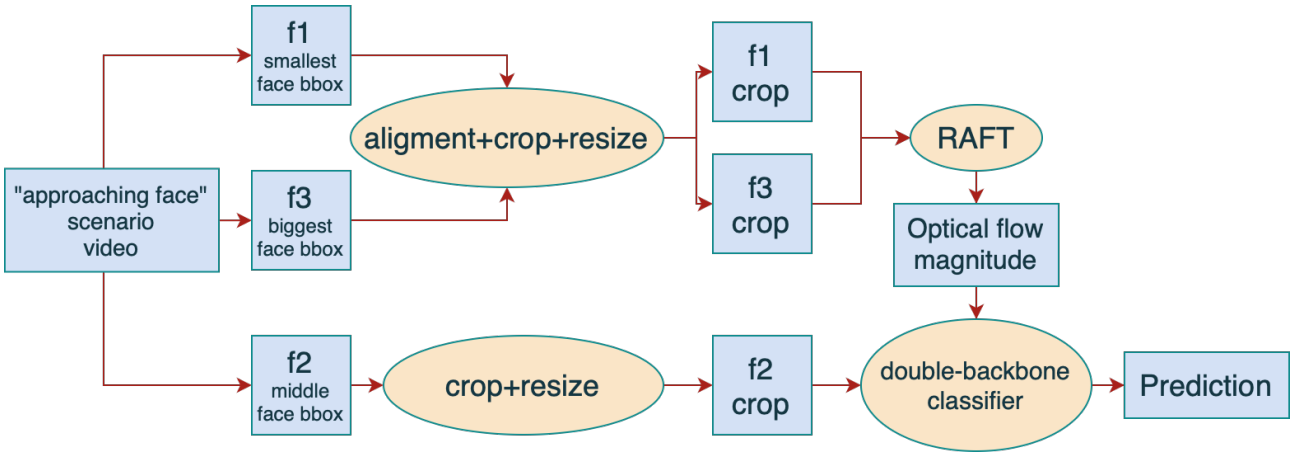

Figure 1. Proposed cooperative liveness detection pipeline

Another approach to the video-based liveness detection is cooperative one, which remains relatively underexplored in literature, with most methods relying on simple interactions, like sentence pronunciation (Kollreider et al., 2007) or facial expression changes (Ming et al., 2018). Such approaches increase the time of verification, but significantly improve the accuracy.

Our method builds upon these foundations while introducing key innovations. Like (Muhammad et al., 2022), we employ stabilization, but enhance it with neural key-point detection for improved alignment. Similar to (Parkin and Grinchuk, 2020), we utilize optical flow, but specifically optimize it for our unique cooperative scenario where controlled face movement provides crucial volumetric information unavailable in passive approaches. Furthermore, we combine optical flow analysis with RGB frame processing to capture both dynamic and static features, improving robustness across diverse attack types from simple printouts to sophisticated 3D masks.

3. Proposed Method

Our pipeline (Figure 1) collects and processes cooperative videos through five stages:

1. **Input frames capturing**: The first (f1), the middle (f2), and the last (f3) frames are extracted according to the cooperative scenario (details in Section 3.1);
2. **Frame preprocessing**: The frames f1, f2, and f3 are aligned and normalized (details in Section 3.2).
3. **Optical flow computation**: We calculate optical flow between the f1 and the f3 using the pretrained RAFT detector (Teed and Deng, 2020).
4. **Optical flow processing**: The magnitude is calculated by the predicted optical flow and then clipped (details in Section 3.3).
5. **Classification**: The processed optical flow and the RGB frame f2 are fed to the neural classifier for final prediction (details in Section 3.4).

3.1 Cooperative scenario

The system guides users through the standardized recording protocol visualizing recorded video on the display (Figure 2). A real-time face detector displays the current face bounding box (blue square). The user first has to align the face with a red reference square (50% frame height), then slowly move it closer to the camera until the face fills an enlarged target square (75% frame height). If the face moves back from the camera or disappears, the recording starts from the beginning. This yields three key frames (f1, f2, f3) with relative face heights of 0.500, 0.625, and 0.750 respectively, that are extracted from the recorded video.

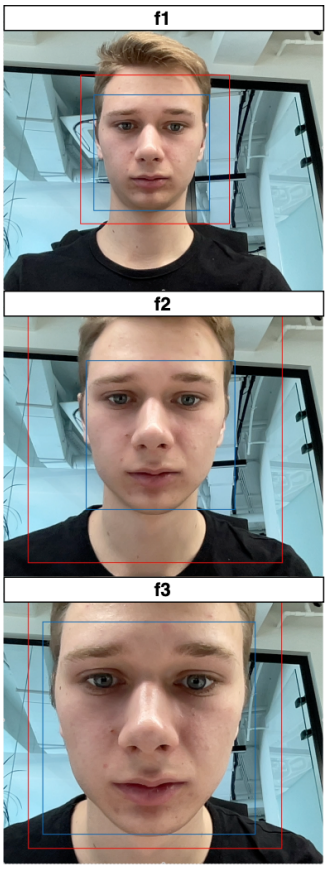

Figure 2. Proposed cooperative scenario

3.2 Input preprocessing

In order to make optical flow detector able to precisely extract facial movements between the first and the last frames of the

video with significant face size variation, we apply specialized preprocessing (Figure 3):

1. Detect facial key-points using a neural detector;
2. Transform f1 and f3 frames through shift and rotation operations to center and align their key-points;
3. Crop the face region with 10% margin around the bounding box;
4. Resize all crops to 256×256 pixels.

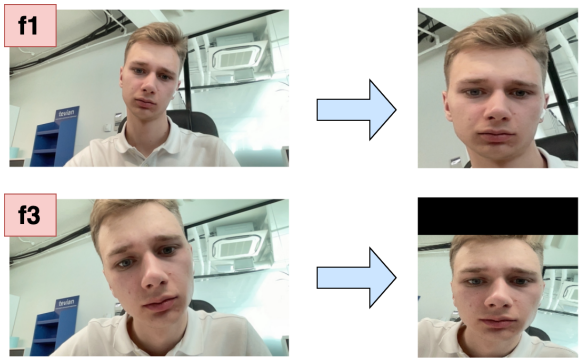

Figure 3. Preprocessing of the f1 and the f3 frames before optical flow computation

The f2 frame undergoes identical processing except for the alignment step.

3.3 Optical flow

We compute optical flow between preprocessed f1 and f3 frames using pretrained RAFT (Teed and Deng, 2020). This reveals distinct patterns for different attack types (Figure 4):

Flat static spoofs (print photos / screen photos). Such fakes' magnitudes show near-zero flow magnitude value in square region around the face (Figure 4c-d).

Flat masks. Such fakes' magnitudes show close to zero flow magnitudes in a face region and high values on the background (Figure 4, (b)).

Real faces and dynamic video replays. Optical flow magnitudes are similar to masks' ones, but demonstrates 3D face patterns and more smooth edges between head and background (Figure 4, (a)).

Since background information is non-informative, we clip magnitude values exceeding 20% of the crop size (51.2 pixels for 256×256 inputs).

3.4 Classification model

For evaluation, we employ a dual-backbone ResNet18 architecture with fully connected fusion layer (Figure 5). The first ”flow backbone” processes the clipped optical flow magnitude between f1 and f3. The second ”RGB backbone” processes the preprocessed f2 frame.

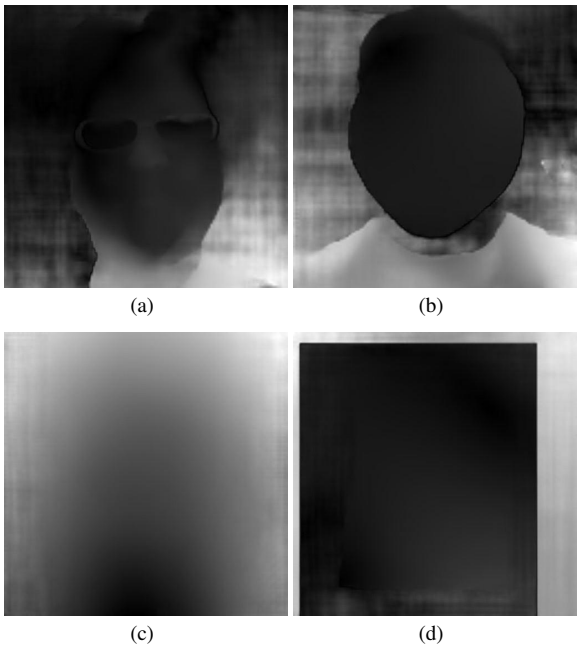

Figure 4. Optical flow magnitude examples: (a) Real face (b) Printed mask attack (c) Screen photo attack (d) Printed photo attack

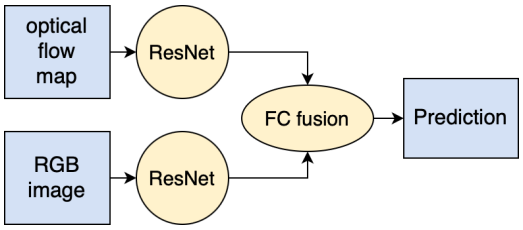

Figure 5. Double-backbone architecture for liveness classification by optical flow map and RGB image

3.5 Training details

For proposed method evaluation both backbones are trained simultaneously. Though in practice, the amount of available data for single shot liveness classification is disproportionately larger than can be collected for the chosen ”approaching face” scenario. That is why RGB backbone can be separately trained on larger single-shot datasets.

In order to expand the training set we employ several optical flow-specific augmentations:

- **Random frame**: Randomly select f1 from the top 10% frames with smallest faces and f3 from the top 10% frames with largest faces;
- **Multi resolution**: Compute flows at resolutions from 192x192 to 320x320 and then resize to 256×256;
- **Perspective augmentation**: Apply random perspective transforms to the f1 and f3 frames before preprocessing to improve robustness to head rotations.

4. Data and Experiments

4.1 Dataset

Since our proposed method relies on a novel ”approaching face” capture scenario, existing public datasets are unsuitable for training and evaluation, as they do not adhere to this specific protocol. To address this gap, we collected a dedicated private dataset. All samples were captured using diverse devices and lighting conditions, with recordings contributed by multiple participants. The dataset strictly follows the ”approaching face” scenario, containing videos of both genuine faces and various types of spoofs (e.g., printed photos, screen replays). Detailed sample statistics are provided in Table 1. The dataset comprises the following subsets:

- **Real.** Set of images with real human faces;
- **Screen Photos.** The subsample shot with photographs of faces demonstrated on various displays;
- **Printed Photos.** The subsample shot with printed photographs;
- **Printed Masks.** The subsample shot with printed full-face masks and with cut-out facial regions;
- **Dynamic Videos.** The subsample shot with prerecorded videos, filmed according to ”approaching face” scenario;
- **Static Videos.** The subsample shot with prerecorded videos with almost static faces;

Videos type	train split	test split
Real	1860	291
Screen Photos	4623	217
Printed Photos	1679	85
Printed Masks	1699	93
Dynamic Videos	870	170
Static Videos	765	218

Table 1. Private dataset statistics (number of samples).

4.2 Evaluation metrics

We measured performance using the ROC AUC metric, computed separately for each spoofing subset (class 0) against the real-face subset (class 1).

4.3 RAFT parameters selection

The optical flow detector’s resolution and number of refinement iterations significantly impact both computational efficiency and face volume information extraction. As the optical flow computation is the bottleneck of the proposed pipeline, we tried to speed it up by using the input data of the smaller resolution and by shortening the number of refinement iterations. For further experiments 256x256 resolution and 3 refinement iterations were fixed, according to experiments shown on Figures 6 and 7.

All the time measurements were conducted in a single thread mode on i5-11600K CPU.

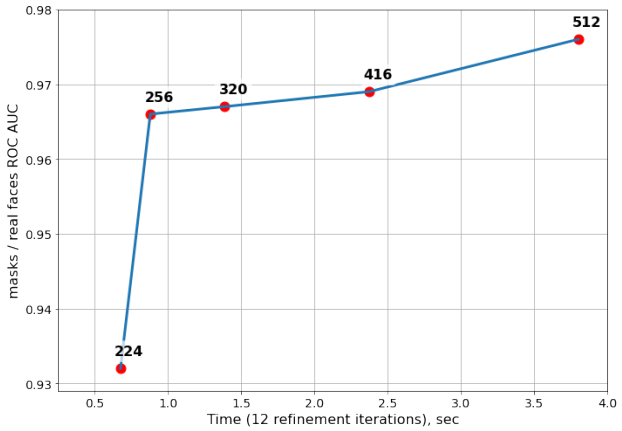

Figure 6. Selection of the OF resolution

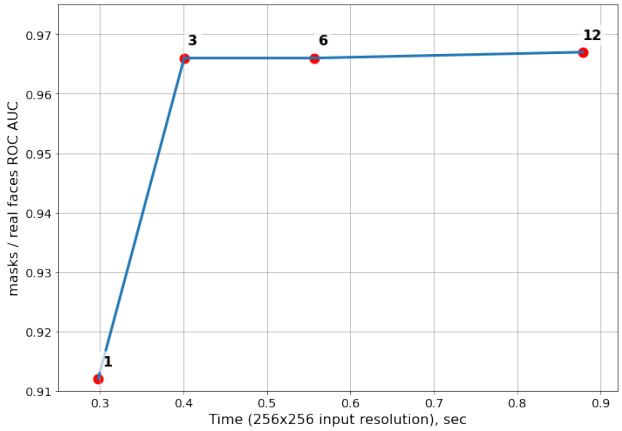

Figure 7. Selection of the RAFT refinement iterations number

4.4 Ablation study

As the baseline model we use the single-ResNet18 classifier that takes optical flow map of size 256x256 as an input. In all our experiments, optical flow map is computed using 3 refinement iterations of RAFT.

4.4.1 Optical flow processing Several options of optical flow processing were considered:

- **Raw optical flows**. Two-channel input with X/Y-axis pixel displacements;
- **Optical flow magnitudes**. Single-channel input with displacement distances for each pixel: $magnitude_{ij} = \sqrt{(x\_shift_{ij} - y\_shift_{ij})^2}$, where i, j the coordinates of a pixel;
- **Clipped optical flow magnitudes**. The magnitude value which are greater than 20% of the input crop side are clipped to $inp\_size * 0.2$.

The table 2 shows that the ROC AUC of the Dynamic Videos subsample is significantly lower than the others. This is due to the fact that the such spoofs’ optical flows are similar with the real faces’ ones.

In further experiments we use cropped optical flow magnitudes as the input of the classifier.

OF preprocessing type	Screen	Printed	Masks	Dynamic Videos	Static Videos
OF	0.992	0.990	0.953	0.668	0.958
OF magnitude	0.991	0.991	0.955	0.670	0.961
clipped OF magnitude	**0.994**	**0.993**	**0.965**	**0.784**	**0.965**

Table 2. Comparison of optical flow preprocessing (ROC AUC)

Augmentations	Screen	Printed	Masks	Dynamic Videos	Static Videos
no augmentations	0.994	0.993	0.965	0.784	0.965
random frame	**0.998**	**0.997**	0.972	0.666	0.972
random frame + multi resolution	**0.998**	0.996	**0.972**	**0.675**	**0.972**

Table 3. Comparison of optical flow augmentations (ROC AUC)

Architecture	Screen	Printed	Masks	Dynamic Videos	Static Videos
OF only; ResNet18	0.998	0.996	0.972	0.675	0.972
OF + RGB (stacked); single ResNet18	0.996	0.994	0.985	0.991	0.998
OF + RGB (separately); double ResNet18	**0.999**	**1.000**	**0.993**	**0.994**	**0.999**

Table 4. Comparison of architectures (ROC AUC)

Modification	Screen	Printed	Masks	Dynamic Videos	Static Videos
None	0.999	1.000	0.993	0.994	0.999
perspective augmentations	**1.000**	**1.000**	**0.994**	**0.996**	**0.999**
perspective augmentations + stabilization	**1.000**	**1.000**	**0.994**	**0.996**	**0.999**

Table 5. Modifications for increacing model stability on samples with head rotations (ROC AUC)

Blurriness level	Screen	Printed	Masks	Dynamic Videos	Static Videos
no blur	1.000	1.000	0.994	0.996	0.999
low	0.997	0.996	0.992	0.989	0.999
medium	0.990	0.981	0.939	0.966	0.984
high	0.937	0.883	0.833	0.846	0.857

Table 6. Evaluation on the blurred test (ROC AUC)

Aggregation	Screen	Printed	Masks	Dynamic Videos	Static Videos
single shot	0.983	0.963	0.970	0.991	0.985
stabilized average	0.999	0.980	0.922	1.000	**1.000**
rank pool + OF	0.993	0.982	0.904	0.921	0.976
proposed method	**1.000**	**1.000**	**0.994**	**0.996**	0.999

Table 7. Aggregation types comparison (ROC AUC)

4.4.2 Optical flow augmentations To extend train set of optical flows we use augmentations, discussed in 3.5.

The table 3 shows that the **random frame** augmentation promoted significant ROC AUC growth, while the **multi resolution** did not affected metrics. However, for overall classifier stability we use both augmentations in further experiments and final pipeline.

4.4.3 RGB input In order to make the model extract textural information from input frames, we modified the architecture to make it able to receive an RGB image along with an optical flow. These are two proposed architecture variants that we evaluated:

- **OF + RGB (stacked); single ResNet18**. Input is a concatenation of the RGB image and the optical flow magnitude;
- **OF + RGB (separately); double ResNet18**. The first backbone processes an optical flow magnitude, the second one processes an RGB image.

The experiments, described in the table 4 shows that the model ability to extract textural information significantly increases ROC AUC metric for all subsets, especially for the Dynamic Videos. In the further experiments we use the dual-ResNet18 architecture as it outperformed other models on all data subsets.

4.4.4 Head rotations To increase the model stability to head rotations between the frames f1 and f3 the following measures have been taken:

- **Face stabilization by facial key-points**. This step in input preprocessing makes model stable to roll rotations;
- **Random perspective transformations**. This augmentation is applied to f1 and f3 images before preprocessing, making model stable to pitch and yaw rotations.

The table 5 shows that the ROC AUC metric does not affected by the discussed changes. That happened, because the current test split does not contain samples with strong head rotations. However, manual tests shows that the pipeline becomes stable to all types of head rotations after adding proposed modifications to the final pipeline.

4.4.5 Blur We checked the stability of the proposed pipeline on blurred images. We used Gaussian Blur function from OpenCV to create images with blur. In our experiments we used

three levels of blurriness: low (kernel size is 1% of the mean f1 face crops hight), medium (kernel size is 6% of the mean f1 face crops hight) and high (kernel size is 12% of the mean f1 face crops hight) (Table 6).

4.5 Comparison with other aggregation methods

We also adapted ideas from (Muhammad et al., 2022) and (Parkin and Grinchuk, 2020) to the proposed cooperative scenario to make a comparison with developed pipeline (Tables 7, 8). These methods were chosen for the comparison as we found them profitable to use with ”approaching face” scenario. Here are the details of the methods adaptation:

4.5.1 Classification by stabilized average frame The idea was taken from (Muhammad et al., 2022). In order to improve face stabilization we use the neural facial key-points detector. The conducted experiments showed that the best algorithm quality is achieved with using 5 frames for computing input video aggregation. We used ResNet18 model for classification.

4.5.2 Classification by optical flow and rank pooling outputs The idea was taken from (Parkin and Grinchuk, 2020). In the adapted pipeline we apply rank pooling to the input videos twice with the same parameters as in the original paper. Also we compute optical flow between the f1 and the f3 frames with RAFT as in the proposed method. The received features are passed to three ResNet18 backbones with fully connected fusion layer as in the proposed method.

Also we compared proposed pipelines with the single-shot ResNet18 model, that was evaluated on the f2 frames.

Method	sec.
single-shot	0.05
stabilized average	0.05
rank pool + OF	0.75
proposed method	0.55

Table 8. Runtime performance comparison (without preprocessing)

5. Conclusion

In this work, we proposed the novel cooperative video-based face liveness detection method that leverages optical flow analysis under a controlled ”approaching face” scenario. Our approach combines stabilized facial key-point alignment with neural optical flow estimation to effectively distinguish between real faces and various types of spoofing attacks, including printed photos, screen displays, masks, and video replays.

References

Fernando, B., Gavves, E., Oramas M., J., Ghodrati, A., Tuytelaars, T., 2017. Rank Pooling for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 773–787. http://dx.doi.org/10.1109/TPAMI.2016.2558148.

Jee, H.-K., Jung, S., Yoo, J.-H., 2006. Liveness Detection for Embedded Face Recognition System. *International Journal of Biological and Medical Sciences*, 1, 235-238.

Kollreider, K., Fronthaler, H., Isaac, M., Bigun, J., 2007. Real-Time Face Detection and Motion Analysis With Application in “Liveness” Assessment. *Information Forensics and Security, IEEE Transactions on*, 2, 548 - 558.

Liang, Y.-C., Qiu, M.-X., Lai, S.-H., 2024. Fiqa-fas: Face image quality assessment based face anti-spoofing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1462–1470.

Marais, M., Brown, D., Connan, J., Boby, A., 2023. Facial liveness and anti-spoofing detection using vision transformers. *Proc. Southern Afr. Telecommun. Netw. Appl. Conf.(SATNAC)*, 1–6.

Ming, Z., Chazalon, J., Luqman, M. M., Visani, M., Burie, J.-C., 2018. Facelivenet: End-to-end networks combining face verification with interactive facial expression-based liveness detection. *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 3507–3512.

Ming, Z., Yu, Z., Al-Ghadi, M., Visani, M., MuzzamilLuqman, M., Burie, J.-C., 2022. Vitranspad: Video transformer using convolution and self-attention for face presentation attack detection.

Muhammad, U., Yu, Z., Komulainen, J., 2022. Self-supervised 2d face presentation attack detection via temporal sequence sampling. *Pattern Recognition Letters*, 156, 15–22.

Nikitin, M. Y., Konushin, V. S., Konushin, A. S., 2019. Face anti-spoofing with joint spoofing medium detection and eye blinking analysis. , 43(4), 618–626.

Parkin, A., Grinchuk, O., 2020. Creating artificial modalities to solve rgb liveness.

Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 402–419.