Using Synthesized Facial Views for Active Face Recognition

Efstratios Kakaletsis* and Nikos Nikolaidis

Department of Informatics, Artificial Intelligence & Information Analysis Laboratory, Aristotle University of Thessaloniki, Thessaloniki, GR-54124, Greece.

*Corresponding author(s). E-mail(s): ekakalets@csd.auth.gr;
Contributing authors: nnik@csd.auth.gr;

Abstract

Active perception / vision exploits the ability of robots to interact with their environment, for example move in space, towards increasing the quantity or quality of information obtained through their sensors and, thus, improving their performance in various perception tasks. Active face recognition is largely understudied in recent literature. Attempting to tackle this situation, in this paper, we propose an active approach that utilizes facial views produced by photorealistic facial image rendering. Essentially, the robot that performs the recognition selects the best among a number of candidate movements around the person of interest by simulating their results through view synthesis. This is accomplished by feeding the robot's face recognizer with a real world facial image acquired in the current position, generating synthesized views that differ by $\pm\boldsymbol{\theta}^{\circ}$ from the current view and deciding, based on the confidence of the recognizer, whether to stay in place or move to the position that corresponds to one of the two synthesized views, in order to acquire a new real image with its sensor. Experimental results in three datasets verify the superior performance of the proposed method compared to the respective "static" approach, approaches based on the same face recognizer that involve synthetic face frontalization and synthesized views, random direction robot movement, robot movement towards a frontal location based on view angle estimation, as well as a state of the art active method. Results from a proof of concept simulation in a robotic simulator are also provided.

Keywords: active vision ; active face recognition ; synthesized facial views; photorealistic facial synthesis

1 Introduction

In recent years, the robotics and vision communities have started researching more thoroughly the field of active vision / perception and exploration. Active perception methods try to obtain more, or better quality, information from the environment by actively choosing from where and how to observe it using a camera (or other sensors), in order to accomplish more effectively tasks such as 3D reconstruction [1, 2], [3], [4], [5] or object recognition [6], [7]. This could be achieved, for example, by moving a camera-equipped mobile robot, e.g. a wheeled robot or a UAV, in positions which offer different (and hopefully better) views of the object of interest. Although active 3D object reconstruction has attracted considerable interest, mainly towards solving the "next-best-view" problem (i.e. choosing the next viewing position in order to to obtain a detailed and complete 3D object model), active approaches for recognition tasks, especially for face recognition, are less frequent in the literature. Deep Learning has lately dominated face recognition research due to the superior performance achieved. However the vast majority of recognition methods adopt a static approach i.e., an approach that is based on an image acquired from a specific viewpoint, even in setups where an active approach could have been used. Indeed, face recognition can be combined with an active approach for controlling the movement of a camera-equipped robot towards capturing the face from more informative views and thus obtaining more robust results, at the expense of energy consumption and additional time needed. Synthesized views of faces, whose images were acquired through a camera, can be used for robot movement guidance in an active face recognition setup. Instead of having the robot move in a physical way for capturing a novel (and better) view, one can use a synthesized view as an aid towards choosing a new viewpoint and improving recognition through an acquisition procedure.

In this paper, we propose an active face recognition approach that utilizes facial views synthesized by photorealistic facial image rendering. Essentially, the camera-equipped robot that performs the recognition selects the best among a number of candidate physical movements around the face of interest by simulating their results through view synthesis. In other words, once the robot (that is at a certain location with respect to the subject) acquires an image, it feeds the face recognizer with this image as well as with synthesized views that differ by $\pm\theta^{\circ}$ from the current view. Subsequently, it either stays in the current position or moves to the position that corresponds to one of the two synthesized views. The respective decision is based on the confidence of the three recognitions (on the real and the two synthesized views). In case of a "move" decision, it proceeds in acquiring a "real" image from its new location. The procedure repeats in the same manner, for this location, for one or more

steps. Using synthesized facial views facilitates decision-making by providing estimates of what is to be expected (in terms of recognition accuracy) in a new robot position. The proposed method involves a face recognizer that is trained to recognize frontal or nearly frontal faces, while having to deal with input facial images obtained from an arbitrary view point. This fact makes recognition challenging, but at the same time more easily applicable in a real-world scenario, since it does not require the existence of facial images acquired from different viewpoints in order to train a view-independent face recognizer.

The remainder of this paper is organized as follows. In Section II related work is presented, whereas in Section III we describe the details of the proposed method. In Section IV experiments conducted to measure the algorithm's performance are presented. Finally, Section V provides a discussion and conclusions.

2 Related Work

2.1 Active Computer Vision

A few recent active approaches for tasks such as object detection, recognition, 3-D reconstruction and manipulation are presented in this section. Additional methods can be found in the review paper [8] that deals in particular with the problem of view planning in robot active vision.

In [9], a robotic arm equipped with a depth camera captures information for a scene from several poses, towards understanding the environment and performing multiple object detection. Boundary Representation Models (B-Reps) are used to represent the objects. The world representation is initialized and, after generating a first set of object detection hypotheses, the approach tries to perform exploration in order to generate new hypotheses or validate existing ones. This is accomplished by finding regions of interest (regions to be inspected) and suitable new views, acquired by appropriate poses of the arm. A proof of concept using a KUKA LWR 4 arm is provided. As expected, the object recognition rate increases as the number of views increases.

In [10] the authors deal with the problem of reconstructing a scene while also identifying the objects in it using 3D scans and a dataset of 3D shapes. Towards this end, a 3D attention model is developed that selects the best views to scan from, as well as the most informative regions within in each view, so as to achieve object recognition. The region-level attention mechanism generates features which are fairly robust against occlusion. Temporal dependencies among consecutive views are encoded with deep recurrent networks.

A new approach, called 3D ShapeNets, for representing a 3D shape as a probability distribution of binary variables on a voxel grid, using a Convolutional Deep Belief Network is proposed in [11]. This representation supports joint object recognition and shape completion from depth maps and enables active object recognition through view planning. The model, learns the distribution of 3D shapes from different object categories and various poses from raw CAD data, while also discovering hierarchical compositional part representations.

Moreover, in [12], the authors present a novel methodology for optimizing a robot's vision sensor viewpoint and apply it in the tasks of object recognition and manipulation (grasping synthesis) in unstructured environments. The algorithm uses extremum seeking control (ESC), which utilizes a task success criterion in a continuous optimization loop. In the case of object recognition, an image is captured by the robot's camera and supplied to the recognition algorithm. The algorithm generates a success rate value (probability of recognizing an object) that forms the main component of the objective function, which is to be maximized by the neural-network based ESC algorithm, towards generating velocity commands for the robot camera. The camera moves on a sphere (viewsphere) around the object, i.e., it points to the object all the time while keeping the distance fixed. The algorithm requires neither a task model nor training on offline image data for viewpoint optimization and is shown to be robust to occlusions.

In [13] another active vision-based object recognition approach is presented, among other contributions. More specifically, a CNN-based approach is described that allows object recognition over arbitrary camera trajectories, (which generate multi-view image sequences) without requiring explicit training over the potentially infinite number of camera paths and lengths. This is done by decomposing an image sequence into a set of image pairs, classifying each pair independently, and then learning an object classifier by weighting the contribution of each pair. The method is then extended to the next-best-view problem in an active recognition framework. This is accomplished by training a second CNN to map from an observed image to the next viewpoint and incorporating it into a trajectory optimisation task.

In [14] a method for active object recognition that involves a deep CNN for the simultaneous prediction of the object label and the next action to be performed by the sensor so as to improve recognition performance is presented. The task is treated as a reinforcement learning problem and a generative model of object similarities is embedded in the network for encoding the state of the system. Other, older, active object recognition approaches, are reviewed in [15–17].

[1] deals with the problem of active object reconstruction. In there, a next-best-view planning scheme based on supervised deep learning is proposed. A properly trained three-dimensional convolutional neural network (3D-CNN) is used to predict the next-best-view position, given the current view.

Finally, in [18] a viewpoint planning strategy for 3D reconstruction with application in the reconstruction of blades is presented. The algorithm focuses on controlling surface overlap for the various views so as to allow for successful registration. OctoMaps are used towards this end and the method is tested in both simulation and real blade reconstruction.

2.2 Active Face Recognition

Despite the fact that active object recognition has attracted considerable interest in the computer vision and robotics communities, active face recognition

has been scarcely studied. Such a simple method is described in [6] and comprises of a neural network-based face recognizer along with a decision making controller that decides for the viewpoint changes. More specifically, a pre-trained VGG-Face CNN is used by the recognition module in order to extract facial image features and it is combined with a nearest-neighbor identity recognition criterion. The simple controller module can make three different decisions based on the uncertainty of the current result (i.e., the distance d between the input image and the closest image in the database of known persons): a) recognize the individual, if d is below a threshold t_1 b) disregard the individual as unknown, if d is above a threshold t_2 or c) reassess the subject by moving to a different viewpoint, if $t_1 < d < t_2$. The direction towards which the movement shall be performed in order to increase the probability of correct recognition is not studied by the authors.

The authors in [7] propose a deep learning-based active perception method for embedding-based face recognition and examine its behavior on a real multi-view face image dataset. The proposed approach can simultaneously extract discriminative embeddings, as well as predict the action that the robot must take (stay in place, move left or right by a certain amount, on a circle centered at the person) in order to get a more discriminative view.

2.3 Multi-view Facial Image Synthesis

A significant number of techniques for synthesizing facial images in novel views appeared in the last years since such images can have a number of applications, e.g., in improving face recognition accuracy. For example, since profile faces usually provide inferior recognition results compared to frontal faces, generative adversarial networks (GANs) based methods for the frontalization of profile facial images [19] or generation of other facial views [20] have been proposed for improving face recognition results.

A method for the generation of frontal views from any input view that utilizes a novel generative adversarial architecture called the Attention Selective Network (ASN) is described in [21]. Towards improving single-sample face recognition by both generating additional samples and eliminating the influence of external factors (illumination, pose), [22] presents an end-to-end network for the estimation of intrinsic properties of a facial image with recovery of albedo UV map and 3D facial shape. In [23], a facial image rendering technique is used both in the training and testing stages of a face recognition approach.

A method that produces photorealistic facial image views is described in [24]. The basic idea of this approach is that rotating faces in the 3D space and re-rendering them to the 2D plane can serve as a strong self-supervision. A 3D head model (obtained by utilizing the 3D-fitting network 3DDFA [25–27]), accompanied by the projected facial texture of a single view, is being rotated and multi-view images of the face are rendered using the Neural 3D Differential Renderer [28] along with 2D-to-3D style transfer and image-to-image translation with GANs to fill in invisible parts. This last state-of-the-art

method was selected due to its robustness and photorealistic quality for the generation of the synthetic facial images required by the method proposed in this paper.

Although facial view synthesis can improve face recognition performance, active perception methods can be expected to provide better results, in cases where acquisition of additional "real' facial views is possible due to the existence of e.g. a wheeled robot.

3 Proposed Active Face Recognition Algorithm

3.1 Face Recognition

Let us denote as database subset G a set of training facial images for the persons that shall be recognized. Similarly, the facial images to feed the face recognizer are included in the query (test) set T. The face recognition library face.evoLVe [29] which contains many state-of-the-art deep face recognition models, is used. More specifically, an implementation of a certain face recognition approach of face.evoLVe from the OpenDr Toolkit [30, 31] was used. IR-50 (50 layers) [32] trained on MS-CELEB-1M using an ArcFace [33] loss head was used as the 512-dimensional feature extraction backbone.

For the database subset G, face detection, facial landmark extraction and face alignment was based on the face.evoLVe module that is based on MTCNN [34], whereas for the query images in T, these processing steps were based on RetinaFace [35, 36]. Face recognition is performed by a nearest-neighbor classifier that uses Euclidean distance in the 512-dimensional feature space to find the database facial image that best matches the query image.

Face recognition confidence $FRC \in [0, 1]$, is also evaluated based on the distance between the input query image and the nearest image in the database G. The FRC is given by the following formula:

$$FRC = 1 - \frac{distance}{threshold} \tag{1}$$

where *distance* is the euclidean distance of query facial image from the nearest neighbor image in the database G and *threshold* is the optimal threshold found by running a pairwise matching experiment on LFW [37].

3.2 Active Face Recognition Using Synthesized Views

The proposed active face recognition algorithm uses the face recognition confidence FRC and facial images synthesized for view angles around the current robot view, in order to select the next robot movement, towards performing a successful recognition. Starting from an initial position, the robot can take one of the following three decisions: stay at the current position, move by θ° to the right or move by θ° to the left, in order to acquire a new image. Depending on the achieved recognition confidence, one or more additional movements, towards the same direction as the first one, might be decided.

Algorithm 1 Active Face Recognition Algorithm on Pseudocode

Input: I_r, $threshold$, θ°
Result: $Person_{ID}(I_r)$

$\alpha = Estimate\_View\_Angle(I_r)$
$I_s^- = Render(\alpha - \theta^\circ, I_r)$
$I_s^+ = Render(\alpha + \theta^\circ, I_r)$
$I = \underset{x \in \{I_r, I_s^-, I_s^+\}}{argmax}(FRC(x))$
if $I = I_r$ **then**
    $I_{ID} = I_r$
    go to 32
else
    if $I = I_s^+$ **then**
        $\theta_{incr} = +\theta^\circ$
    else
        $\theta_{incr} = -\theta^\circ$
    end if
end if

$I_r^{1step} = Move\_and\_Capture(\alpha + \theta_{incr})$
if $FRC(I_r^{1step}) > threshold$ **then**
    $I_{ID} = \underset{x \in \{I_r, I_r^{1step}\}}{argmax}(FRC(x))$
    go to 32
else
    $I_s^{2step} = Render(\alpha + 2 * \theta_{incr}, I_r^{1step})$
    if $FRC(I_s^{2step}) < FRC(I_r^{1step})$ **then**
        $I_{ID} = \underset{x \in \{I_r, I_r^{1step}\}}{argmax}(FRC(x))$
        go to 32
    else
        $I_r^{2step} = Move\_and\_Capture(\alpha + 2 * \theta_{incr})$
        $I_{ID} = \underset{x \in \{I_r, I_r^{1step}, I_r^{2step}\}}{argmax}(FRC(x))$
        go to 32
    end if
end if

$Person_{ID}(I_r) = Recognize(I_{ID})$

More specifically, given a facial query image I_r (subscript r stands for real), captured by the robot camera at the robot starting position, the face synthesis algorithm [24] is utilized to estimate the robot view angle α and then render/generate facial views in 2 different view angles i.e. $-\theta^\circ$ and $+\theta^\circ$ in pan with respect to the pan of I_r (and the same tilt as I_r). These two images are

denoted by I_s^- and I_s^+ respectively (subscript s stands for synthetic). Then, the face recognizer is fed with these three images I_r, I_s^-, I_s^+ (one real, two synthetic ones). Depending on the image that obtained the biggest face recognition confidence FRC, the robot stays in its current position (if FRC was maximum in I_r) or physically moves $-\theta^\circ$ (or $+\theta^\circ$) (if FRC was maximum in I_s^- (or I_s^+)) and acquires through its camera a new real image I_r^- (or I_r^+). If a "stay" decision was taken, the algorithm outputs the ID of the person it recognized in I_r and terminates. If the robot moved, face recognition is performed again in I_r^- (or I_r^+) and the obtained FRC is compared to an experimentally evaluated threshold t. In case a high enough confidence was observed, the algorithm outputs the ID of the person it recognized in I_r^- (or I_r^+) and terminates. If not, it tries additional $+\theta^\circ$ steps (movements) in pan, in the same direction as the first step. In more detail, in this second step, it generates/synthesizes a facial view $-\theta^\circ$ (or $+\theta^\circ$) in pan from the current pan value (and the same tilt), denoted as I_s^{--} (or I_s^{++}), and evaluates (by calling the face recogniser) FRC on this synthetic image. If $FRC(I_r^-) > FRC(I_s^{--})$ (or $FRC(I_r^+) > FRC(I_s^{++})$) the algorithm decides that the robot shall stay in its current position, outputs the ID of the person it recognized in I_r^- (or I_r^+) and terminates. Otherwise, the robot physically moves $-\theta^\circ$ ($+\theta^\circ$) from its current position, acquires a new image I_r^{--} (I_r^{++}) and the algorithm outputs the ID of the person it recognized in this image. The procedure can be repeated for a number of additional steps (movements), until the predefined maximum number of steps is reached. The performance of the proposed procedure obviously depends on whether the synthesis algorithm [24] estimates with sufficient accuracy the view angle of the query image I_r and also on whether the synthesized views are of good quality. In order to limit the possibly negative effect of these factors on the performance of the algorithm (e.g. by leading it to move towards the wrong direction), the algorithm does not actually take a decision based on the last real image it has visited but does so based on the real image where it has obtained the maximum FRC value. In more detail, if the algorithm took one step of $-\theta^\circ$, it takes a decision using the real image I given by:

$$I = \underset{x \in \{I_r^-, I_r\}}{argmax} \left(FRC(x)\right) \tag{2}$$

or the equivalent expression that involves I_r^+, I_r, if a step of $+\theta^\circ$ has been taken. Similarly, if two steps of $-\theta^\circ$ each have been performed, the algorithms decides on the person ID using the real image I given by:

$$I = \underset{x \in \{I_r^{--}, I_r^-, I_r\}}{argmax} \left(FRC(x)\right) \tag{3}$$

or the equivalent expression that involves I_r^{++}, I_r^+, I_r, if two steps, of $+\theta^\circ$ each, have been taken. The pseudocode for the proposed method, when two steps are allowed, is presented in algorithm 1.

It should be noted that the actual recognition is always performed on a real image, i.e., an image captured by the robot camera. The synthesized views are only used to aid the robot in deciding whether to move in a new position (and acquire a new image there) or stay in the current position. The rationale behind the proposed approach is that in case the initial robot position is far from a frontal or nearly frontal one, the algorithm will hopefully direct it to move towards a position which is closer to a frontal one. Obviously, the procedure can work, in the same way, for tilt.

4 Performance Evaluation

For the evaluation of the proposed active face recognition approach, a number of experiments were conducted using the HPID dataset [38], the Queen Mary University of London Multi-view Face Dataset (QMUL)[39] and a Synthetic Dataset (SD) [40]. In all three datasets, images of all subjects were divided into two non-overlapping subsets: a database subset G (images that the face recognizer uses to decide upon the ID of the query image through the nearest neighbor classifier) and a query (test) subset T (these are meant to be the images captured by the robot camera in its initial position). This was done by choosing images with different pan ranges for G and T. With this setup we are simulating active recognition where the robot is moving only in the pan direction. Short descriptions of the three datasets are provided below.

4.1 HPID Dataset

The HPID dataset [38] is a head pose image database consisting of 2790 face images of 15 subjects captured by varying the pan and tilt angles from $-90°$ to $+90°$, in $15°$ increments. Two series of images were captured for each person, (93 images in each series).

The database subset G (Figure 1) contains facial images with tilt in angles $[-30°, -15°, 0°, +15°, +30°]$ and pans $[-15°, 0°]$, i.e. only nearly frontal images. The query subset (Figure 2) contains face images with tilts $[-30°, -15°, 0°, +15°, +30°]$ and pans $[-90°, -75°, -60°, -45°, -30°]$. The selection of the range $[-90°... - 30°]$ in pan, instead of the full ($[-90°... - 30°]$ and $[+30°... + 90°]$) semi-circle, in this and the other two datasets, was just for simplicity. Similar results were obtained in experiments involving the full semi-circle.

Synthetic images generated from the "real" query images for use from our algorithm are depicted in Figure 3.

4.2 QMUL Dataset

Queen Mary University of London Multi-view Face Dataset (QMUL) [39] consists of automatically aligned, cropped and normalised face images of 48 persons. Images of 37 persons are in greyscale (dimensions: 100x100 pixels) whereas those of the remaining 11 subjects are in colour and of dimensions

56x56. For each person 133 facial images exist, covering a viewsphere of $-90°...+90°$ in pan and $-30°...+30°$ in tilt in $10°$ increments. For the Database split G, the facial images with pan in angles $[-10°, 0°]$ and tilt in angles $[-30°, ..., +30°]$ were used. The Query split T (test) includes images with pan in angles $[-90°, ..., -20°]$ and tilt in the range $[-30°, ..., +30°]$.

4.3 Synthetic Dataset

The Synthetic Dataset (SD) was generated using Unity's Perception package. It consists of 175422 cropped face images of 33 subjects taken at different environments, lighting conditions, camera distances and angles. In total, the dataset contains images for 8 environments, 4 lighting conditions, 7 camera distances (1m-4m) and 36 camera angles ($0 - 360°$ at $10°$ intervals). A subset of the dataset was used in the experiments. The subset included all 33 subjects in all environments and 1 lighting condition, at a camera distance of 1.0 m. For the Database split G, facial images with pan $[0°, +10°]$ and tilt $0°$ were used. The Query (test) split T included images with pan in the range $[+20°, ..., +90°]$ and tilt $0°$.

Fig. 1 Samples from the database subset G of the HPID dataset, depicting real facial images of a subject with tilt angles $[-30°, -15°, 0°, 15°, 30°]$ and pans $[-15°, 0°, 15°]$.

4.4 Results

Results (in terms of recognition accuracy) are presented in Table 1. The line marked "Static" in this Table presents the result of the static equivalent of our approach, in which only the initial query facial image is used by the same recogniser involved in the active approach. As can be seen, the proposed active method (lines "Proposed (Active), 2 steps" and "Proposed (Active), 4 steps", referring to the cases where the robot can move up to 2 or 4 times from its initial position in $\theta°$ increments) outperforms its static counterpart for both 2 and 4 algorithm steps, at the obvious expense of additional robot movements and time required to perform them. For the HPID and SD datasets the best performance is obtained for 4 steps of the algorithm and the absolute increase of accuracy with respect to the static version is 15.61% and 13.05% respectively, whereas for the QMUL dataset the best performance is obtained for 2 steps (increase of 15.69% compared to the static approach).

Fig. 2 Samples from the query subset T depicting real facial images of a subject with tilts $[-30°, -15°, 0°, 15°, 30°]$ and pans $[-90°, -75°, -60°, -45°, -30°]$.

Fig. 3 Samples of synthetic facial images generated from the query subset T of the HPID dataset. Each row depicts the two synthetic images generated in pan angles $pan - 15°$, $pan + 15°$ from real images with pans $[-75°, -60°, -45°, -30°]$. Each row corresponds to a different tilt value of the real image, in the range $[-30°, -15°, 0°, 15°, 30°]$.

Fig. 4 Samples from the database subset G of the SD dataset, depicting facial images of four subjects with tilt angle $0°$ and pans $[0°, +10°]$.

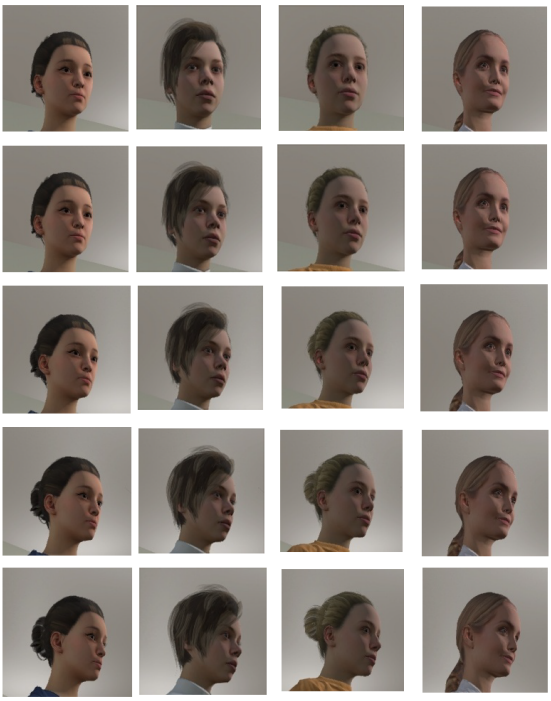

Fig. 5 Samples from the query subset T of SD Dataset depicting facial images of four subjects with tilt 0° and pans $[+20°, +30°, +40°, +50°, +60°]$.

Table 1 Face recognition accuracy results and comparison with the static approach and other variants

Method	HPID[38]	QMUL [39]	Synthetic(SD) [40]
Static *(only Queries)*	72.49 %	69.88%	66.95%
Proposed *(Active)* (2 steps)	82.12%	**85.57%**	68.35 %
Proposed *(Active)* (4 steps)	**88.10%**	82.85%	**80%**
Random direction movement *(2 steps)*	71.31%	72.68%	63.54%
Frontalization by physical movement *(real frontal views)*	74.82%	62.83%	56.33%
Frontalization *(synthetic frontal views)*	80.75%	75.95%	66.10%
Static & Synthetic *(real & synthetic views)* (4 steps)	72.22%	66.35%	62.28%

Table 2 Comparison with [7]

Method	HPID [38]	QMUL [39]	Synthetic (SD) [40]
Proposed *(Active)* (2 steps)	82.88%	82.47%	85.00%
Proposed *(Active)* (4 steps)	**87.78%**	**84.59%**	**88.81%**
[7] *(Active)* (2 steps)	60.96%	69.94%	67.63%
[7] *(Active)* (4 steps)	61.30%	68.11%	70.41%

The proposed approach was also compared to the frontalization approach that is often used in face recognition when the recognizer is trained only on

frontal views. In this case, the facial view synthesis algorithm [24] is used in order to generate a frontal (0° in pan) view from the input (query) image. This image is then provided to the recognizer. The results (line "Frontalization (synthetic frontal views)") show that although frontalization achieves improved performance with respect to the static approach in HPID and QMUL datasets and similar performance in SD, it is superseded by the proposed active approach. Indeed the best achieved results of the proposed approach correspond to an absolute increase in accuracy (with respect to the frontalization approach) of 7.35%, 9.62% and 13.9% for the HPID, QMUL and SD datasets respectively.

One can naturally wonder what is the benefit of introducing an active approach, that involves actual robot movement, over the use of synthetic images only. To answer this question we set up another experiment where for each (real) query image, captured at a view angle α we generate (where possible) 8 synthetic images at angles $\alpha \pm \theta^\circ, ..., \alpha \pm 4\theta^\circ$ around the query and feed them to the recognizer along with the query image. The result with the highest FRC is then adopted as the final decision. Results are presented in line "Static & Synthetic (real & synthetic views) (4 steps)". Obviously this approach is not viable, providing results inferior to those of the static case.

One could also argue that, instead of using the synthesized views as proposed in this paper, it would suffice to estimate the view angle of the robot with respect to the person and instruct it to move directly (i.e., without intermediate steps) to the position that would allow it to obtain a frontal view (0° in pan). However, there are certain difficulties that would make such an approach hard to implement in practice. Indeed, we observed during the experiments that view angle estimates (at least those provided by the view synthesis algorithm used in this paper) although accurate enough for the purposes of view synthesis, are quite far from the ground truth values, thus deeming this approach problematic. Experiments were performed to quantitatively verify this claim. The experimental evaluation was conducted on all three datasets, and involved obtaining the view angle estimate θ° and instructing the robot to physically move by $-\theta^\circ$ and recognise the subject from its -supposedly- frontal new position. The respective recognition accuracy figures are provided in the line "Frontalization by physical movement (real frontal views)" of Table 1. The obtained results show that this approach is significantly inferior to the proposed one and also worse than the frontalization approach that is based on view synthesis. As a matter of fact, this approach is inferior to even the static one in two out of three datasets.

Another set of experiments were conducted in order to prove that the guidance provided by the synthesized views with respect to the direction the robot shall move is beneficial for the proposed algorithm. Towards this end, the proposed approach was compared to a two-step random direction movement approach that was implemented as follows: starting from its initial position, the robot chooses a random direction (positive or negative rotation, i.e., right or left movement) and then performs two θ° steps ($\theta^\circ = 10^\circ$ or 15° depending on

the dataset) towards this direction, capturing the respective (real) images. The decision on the ID of the depicted person is then taken based on the real image (one of the three available) where the maximum FRC value was observed. This approach is similar to the 2 step version of the proposed algorithm, the difference being that the movement direction is not decided on the basis of the utilized synthetic views but is chosen randomly. The recognition accuracy results for this approach are presented in row "Random direction movement (2 steps) of Table 1. By observing this Table, one can notice that results are close to those of the static approach but clearly inferior to those obtained by the 2 step version of the proposed algorithm. This verifies that the guidance provided by the synthetic images is indeed beneficial for the recognition.

Finally, the proposed method was compared to the recent embedding-based active deep face recognition technique [7]. The experimental setup followed in [7] for the HPID datased, was used in all three datasets. More specifically, 75% of the subjects contained in each dataset was used to train the models of [7], while the remaining 25% were used for evaluating the trained models (test set). Since our approach requires no training, only the test set data were utilized in the experiments that involved it. Images in the test set were used to form the Database split G and the Query split T, in the same way (same range of pan and tilt angles) mentioned in Sections 4.1 to 4.3. Results are presented in Table 2. One can observe that the proposed method outperforms the method in [7] in both the 2 and 4 steps setups, achieving (in the 4 steps setup) an absolute increase in accuracy of 26.48%, 16.48% and 18.40% for the HPID, QMUL and SD datasets respectively.

Statistics regarding the steps taken by the proposed approach (4 steps) are presented in Table 3 for the SD dataset. Each row in this Table corresponds to the type of real image that the algorithm reached in its course, i.e., the number of steps it has taken towards the right or the left direction. These types are mentioned in the first column and follow the same naming conventions used in Section 3.2. For example, the row marked I_r^+ includes statistics for cases where the algorithm (robot) moved by $+10°$ from its initial position (the one represented by the input query image). The pan angle increment from the initial position, the number of images and the percentage they represent over the total are presented for each case. The presented statistics show that in 24.34% of the cases the robot decided to stay in its initial position whereas in the remaining 75.66% it moved by $\pm 10°, .., \pm 40°$ (one to four steps). It shall be noted however that the decision on the ID of the depicted person is not necessarily obtained from the last position the robot has visited, due to the fact that the image with the maximum recognition confidence (FRC) is used for this purpose.

The average number of movements that the algorithm instructs the robot to perform can be easily evaluated from statistics such as the ones presented in Table 3. The respective figures are presented in Table 4. Note that in case the robot decides to performs no movement (stay decision) the number of movements is obviously zero. As can be seen, when 4 steps are allowed, the

Table 3 Active Face Recognition Statistics (4 Steps, SD dataset): steps performed by the algorithm.

Image type	Angle	# Images	Percentage
I_r	0°	28	24.34%
I_r^{+}	+10°	5	4.347%
I_r^{++}	+20°	7	6.086%
I_r^{+++}	+30°	5	4.347%
I_r^{++++}	+40°	3	2.608%
I_r^{-}	−10°	52	45.217%
I_r^{--}	−20°	8	6.956%
I_r^{---}	−30°	4	3.478%
I_r^{----}	−40°	3	2.608%
Total	–	115	100%

algorithm instructs the robot to make, on average, from 0.76 to 1.17 movements, a fact that denotes that the time required for active recognition (time for the computations as well as the time for the robot to move) is relatively low and can be further lowered if only 2 steps are allowed.

Table 4 Active Face Recognition Statistics: average number of steps.

Method	**HPID [38]**	**QMUL [39]**	**SD [40]**
Proposed *(Active)* 2 steps	0.82	0.6689	1.14
Proposed *(Active)* 4 steps	**0.89**	**0.7623**	**1.17**

In addition, Table 5 presents statistics regarding the moves that the algorithm (equivalently the robot in a real situation) performs and whether these lead towards a frontal view, i.e., 0° in pan (which is something that might be expected since the recognised is trained on near frontal images) or away from such a view. The statistics for the HPID, show that in most cases (59.6%) the algorithm moves the robot towards a frontal view. However, in another 20.6% of the cases the robot moves away from the frontal position, which indicates that either the estimate for the view angle of the input (query) image provided by the view synthesis algorithm is rather inaccurate or that the generated synthetic views are in some cases of poor quality, causing the algorithm to err with respect to the direction it shall move the robot. A similar behavior can be observed in the SD dataset, whereas in QMUL in the majority of cases 49.35% the algorithm decides to stay in the initial position whereas it moves away from the frontal direction in 45.73% of the cases (Table 5). Despite these issues, the algorithm manages to achieve good results in most cases.

Table 5 Active Face Recognition Statistics: move type (4 steps)

Move Type	**HPID [38]**	**QMUL [39]**	**SD [40]**
towards frontal	**447**(59.67%)	57(4.9%)	**67**(58.26%)
stay	117(15.62%)	**573**(49.35%)	28(24.34%)
away from frontal	155(20.69%)	531(45.73%)	20(17.39%)
total	749	1161	115

4.5 Simulation Results

In order to provide simulation-based evidence that the proposed active vision method is indeed effective, we created a simple simulation environment using the open source and widely used Webots [41],[42] robotic simulator. The environment implements a face recognition scenario that involves a TIAGo mobile manipulator robot[1] and its RGB camera. The TIAGo model[2] provided with the simulator moves in a circle around a static human model and performs face recognition using the proposed active method (2 steps approach). The method involves the same face detector and recogniser used in the experiments performed on the three datasets. The Database split G of the face recogniser contains facial images from 10 subjects with pan $[0°, \pm 15°]$ and tilt $0°$. In the implemented scenario, the robot is placed (initialized) at a random location approximately 2m away form the subject and performs active recognition on the face detected in the frames of its camera.

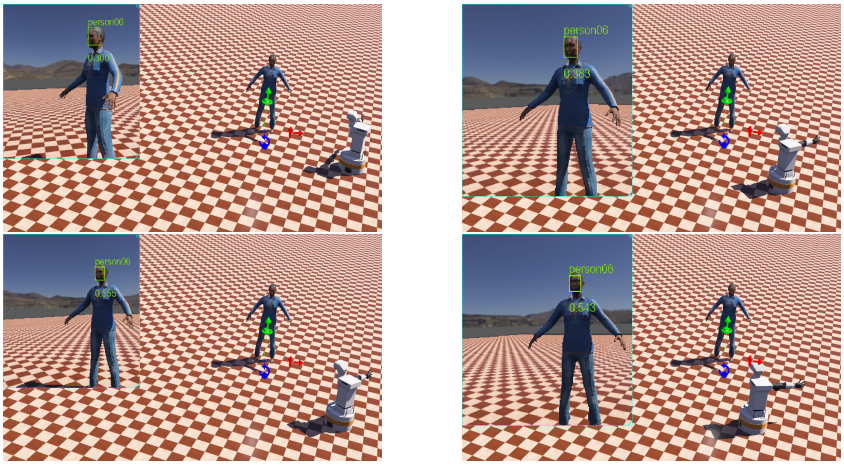

Fig. 6 A simulation in Webots where a TIAGo mobile robot performs active face recognition on person 06 starting from two different initial robot positions (top row) and reaching its final position (respective image at the bottom row). The sub-image at the upper-left corner depicts the robot camera view along with the face detection bounding box, person label and recognition confidence.

[1] https://pal-robotics.com/robots/tiago/
[2] https://cyberbotics.com/doc/guide/tiago-steel

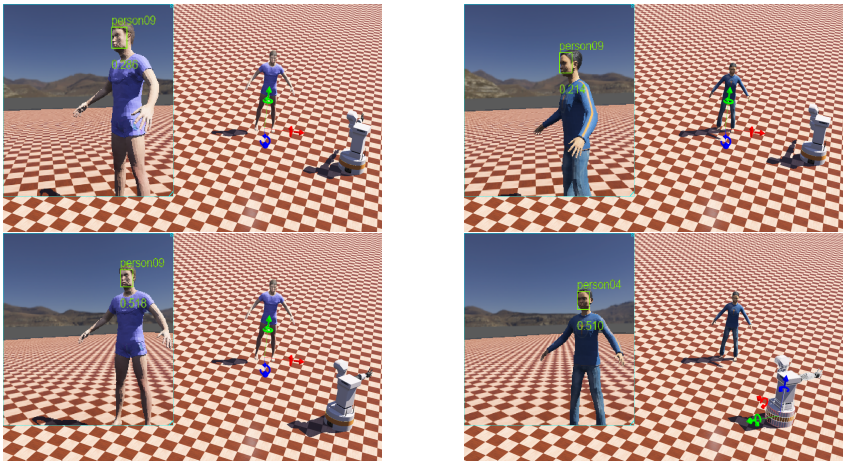

Fig. 7 Additional simulation examples involving persons 09 (left) and 04 (right). Top row: initial robot positions. Bottom row: final robot position. Notice that in the case of person 04 the robot recognizes a different one (person 09) in its initial position, however this is corrected in its final position.

As illustrated in Figures 6 and 7 the robot moves towards more frontal views, increasing the recognition confidence and, in one case (Figure 7 right), changes its decision regarding the person's identity, towards the correct one.

5 Discussion and Conclusions

An active face recognition approach that utilizes facial views produced by facial image synthesis was presented in this paper. The camera-equipped robot that performs the recognition selects the best among a number of candidate physical movements around the person of interest by simulating their results through view synthesis. Experimental results show that the proposed method is superior to both its static version and face recognition that involves synthetically generated images. Moreover, it achieves significantly better results than the method in [7].

It shall be noted that certain assumptions were adopted in this paper and, furthermore, a number of issues were not fully addressed. First of all, the actual control of the robot in order to move in θ° increments around the person is not dealt with, being outside the scope of the paper. However, a rough estimate of the person position with respect to the robot would suffice to enable robot control. Also, it was assumed that the person being recognized remains relatively static during the recognition process, which can be a fair assumption if this process is brief or in situations when the person is siting or lying on a bed. In case the person moves during this process, this shall be taken into account by the algorithm.

Moreover, it was assumed that there are no obstacles in the robot path. If this is not the case, these obstacles shall be detected (e.g. by a depth sensors)

and taken into account when planning the next move. Furthermore, obstacles in the space between the robot and the person might occlude the person for certain robot-person relative positions. However, since the algorithm decides on the person's identity based on the acquired image where the recognizer obtained the largest recognition confidence, it is rather safe to assume that, in most such cases, the algorithm might not face serious problems, even if it has instructed the robot to move in positions where occlusions occur.

Regarding algorithm performance, as mentioned in the previous section, there is a significant number of cases where the algorithm instructs the robot to move in a direction that is not towards a more frontal view. This might be attributed to errors of the view angle estimation and view synthesis algorithms. Using a better algorithm of this type might possibly lead to even bigger improvements with respect to the static approach. Another useful observation is that, giving the robot the freedom to move for additional steps (4 instead of 2) does, in two of the three datasets, significantly improve the recognition accuracy.

In the future, we plan to evaluate the proposed algorithm in larger datasets and extend the Webots simulation in order to investigate some of the issues mentioned above (occlusions, objects that hinder robot motion etc). Comparison of our approach to additional methods is also planned.

Acknowledgments

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871449 (OpenDR). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

References

[1] Miguel Mendoza, J Irving Vasquez-Gomez, Hind Taud, L Enrique Sucar, and Carolina Reta. Supervised learning of the next-best-view for 3D object reconstruction. *Pattern Recognition Letters*, 133:224–231, 2020.

[2] Jeffrey Delmerico, Stefan Isler, Reza Sabzevari, and Davide Scaramuzza. A comparison of volumetric information gain metrics for active 3D object reconstruction. *Autonomous Robots*, 42(2):197–208, 2018.

[3] Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3D reconstruction. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 3477–3484. IEEE, 2016.

[4] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Appearance-based active, monocular, dense reconstruction for micro aerial vehicles. In

Proceedings of Robotics: Science and Systems, Berkeley, USA, July 2014.

[5] J Irving Vasquez-Gomez, David Troncoso, Israel Becerra, Enrique Sucar, and Rafael Murrieta-Cid. Next-best-view regression using a 3d convolutional neural network. *Machine Vision and Applications*, 32(2):1–14, 2021.

[6] Masaki Nakada, Han Wang, and Demetri Terzopoulos. Acfr: Active face recognition using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 35–40, 2017.

[7] Nikolaos Passalis and Anastasios Tefas. Leveraging active perception for improving embedding-based deep face recognition. In *Proceedings of IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2020.

[8] Rui Zeng, Yuhui Wen, Wang Zhao, and Yong-Jin Liu. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 6(3):225–245, 2020.

[9] Dorian Rohner and Dominik Henrich. Using active vision for enhancing an surface-based object recognition approach. In *Proceedings of Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 375–382. IEEE, 2020.

[10] Kai Xu, Yifei Shi, Lintao Zheng, Junyu Zhang, Min Liu, Hui Huang, Hao Su, Daniel Cohen-Or, and Baoquan Chen. 3D attention-driven depth acquisition for object identification. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016.

[11] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015.

[12] Berk Calli, Wouter Caarls, Martijn Wisse, and Pieter P Jonker. Active vision via extremum seeking for robots in unstructured environments: Applications in object recognition and manipulation. *IEEE Transactions on Automation Science and Engineering*, 15(4):1810–1822, 2018.

[13] Edward Johns, Stefan Leutenegger, and Andrew J Davison. Pairwise decomposition of image sequences for active multi-view recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3813–3822, 2016.

[14] Mohsen Malmir, Karan Sikka, Deborah Forster, Ian Fasel, Javier R Movellan, and Garrison W Cottrell. Deep active object recognition by joint label and action prediction. *Computer Vision and Image Understanding*, 156:128–137, 2017.

[15] Shengyong Chen, Youfu Li, and Ngai Ming Kwok. Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research*, 30(11):1343–1377, 2011.

[16] Sumantra Dutta Roy, Santanu Chaudhury, and Subhashis Banerjee. Active recognition through next view planning: a survey. *Pattern Recognition*, 37(3):429–446, 2004.

[17] GCHE de Croon, Ida G Sprinkhuizen-Kuyper, and Eric O Postma. Comparing active vision models. *Image and Vision Computing*, 27(4):374–384, 2009.

[18] Weixing Peng, Yaonan Wang, Zhiqiang Miao, Mingtao Feng, and Yongpeng Tang. Viewpoints planning for active 3-d reconstruction of profiled blades using estimated occupancy probabilities (EOP). *IEEE Transactions on Industrial Electronics*, 68(5):4109–4119, 2020.

[19] Qingyan Duan and Lei Zhang. Look more into occlusion: Realistic face frontalization and recognition with boostgan. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):214 – 228, January 2021.

[20] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision, (ICCV)*, pages 2439–2448, 2017.

[21] Jiashu Liao, Alex Kot, Tanaya Guha, and Victor Sanchez. Attention selective network for face synthesis and pose-invariant face recognition. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages 748–752. IEEE, 2020.

[22] Huan Tu, Gesang Duoji, Qijun Zhao, and Shuang Wu. Improved single sample per person face recognition via enriching intra-variation and invariant features. *Applied Sciences*, 10(2):601, 2020.

[23] Iacopo Masi, Tal Hassner, Anh Tuân Tran, and Gérard Medioni. Rapid synthesis of massive face sets for improved face recognition. In *Proceedings of 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 604–611. IEEE, 2017.

[24] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-Render: Unsupervised photorealistic face rotation from single-view

images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5911–5920, 2020.

[25] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3DDFA. https://github.com/cleardusk/3DDFA, 2018.

[26] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 152–168. Springer International Publishing, 2020.

[27] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):78–92, 2017.

[28] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3907–3916, 2018.

[29] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face. evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*, 2021.

[30] Nikolaos Passalis, Stefania Pedrazzi, Robert Babuska, Wolfram Burgard, Daniel Dias, Francesco Ferro, Moncef Gabbouj, Ole Green, Alexandros Iosifidis, Erdal Kayacan, Jens Kober, Olivier Michel, Nikos Nikolaidis, Paraskevi Nousi, Roel Pieters, Maria Tzelepi, Abhinav Valada, and Anastasios Tefas. OpenDR: An Open Toolkit for Enabling High Performance, Low Footprint Deep Learning for Robotics. *arXiv preprint arXiv:2203.00403*, 2022.

[31] OpenDR: A modular, open and non-proprietary toolkit for core robotic functionalities by harnessing deep learning. https://github.com/opendr-eu/opendr. Accessed: 2022-06-27.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[34] 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, 2016.

[35] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5203–5212, 2020.

[36] OpenDR Face Detection module: RetinaFace. https://github.com/opendr-eu/opendr/blob/master/docs/reference/face-detection-2d-retinaface.md. Accessed: 2022-06-27.

[37] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[38] Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial structures. In *Proceedings of Workshop on Visual Observation of Deictic Cestures*, volume 6, page 7. FGnet (IST–2000–26434) Cambridge, UK, 2004.

[39] Jamie Sherrah and Shaogang Gong. Fusion of perceptual cues for robust tracking of head pose and position. *Pattern Recognition*, 34(8):1565–1572, 2001.

[40] Charalampos Georgiadis. Generation of a synthetic annotated dataset for training and evaluating active perception methods. *BSc Thesis, Aristotle University of Thessaloniki*, 2022, doi : 10.13140/RG.2.2.21002.34248.

[41] Webots. http://www.cyberbotics.com. Open-source Mobile Robot Simulation Software.

[42] O. Michel. Webots: Professional mobile robot simulation. *Journal of Advanced Robotics Systems*, 1(1):39–42, 2004.