U3PT: A New Dataset for Unconstrained 3D Pose Tracking Evaluation

Ngoc-Trung Tran[1,2(✉)], Fakhreddine Ababsa[2], and Maurice Charbit[1]

[1] LTCI-CNRS, Telecom ParisTECH, 37-39, Rue Dareau, 75014 Paris, France
trung-ngoc.tran@telecom-paristech.fr
[2] IBISC, University of Evry, 40, Rue du Pelvoux, 91020 Evry, France

Abstract. 3D pose tracking using monocular cameras is an important topic, which has been receiving a great attention since last decades. It is useful in many domains such as: Video Surveillance, Human-Computer Interface, Biometrics, etc. The problem gets much challenging if occurring, for example, fast motion, out-of-plane rotation, the illumination changes, expression, or occlusions. In the literature, there are some datasets reported for 3D pose tracking evaluation, however, all of them retains simple background, no-expression, slow motion, frontal rotation, or no-occlusion. It is not enough to test advances of in-the-wild tracking. Indeed, collecting accurate ground-truth of 3D pose is difficult because some special devices or sensors are required. In addition, the magnetic sensors usually used for 3D pose ground-truth, is uncomfortable to wear and move because of their wires. In this paper, we propose a new recording system that allows people move more comfortable. We create a new challenging dataset, named U3PT (Unconstrained 3D Pose Tracking). It could be considered as a benchmark to evaluate and compare the robustness and precision of state-of-the-art methods that aims to work in-the-wild. This paper will also present the performances of two well-known state-of-the-art methods compared to our method on face tracking when applied to this database. We have carried out several experiments and have reported advantages and some limitations to be improved in the future.

Keywords: 3D pose tracking · 3D head tracking · Pose estimation · Unconstrained pose tracking · Pose tracking dataset · Synthetic data

1 Introduction

The main goal of head pose tracking is to estimate the 6 Degrees-of-Freedom (DoF) - consists of 3D translation and three axial rotations - of a person's head relative to the camera view. As commonly-use in the literature, we adopt three terms Yaw (or Pan), Pitch (Tilt) and Roll for three axial rotations. The detail related to pose estimation in general could be found in this interesting survey (Murphy-Chutorian and Trivedi [2009]), whereas we just consider the pose estimation via tracking approaches because it is much more accurate and applicable to many applications nowadays.

G. Azzopardi and N. Petkov (Eds.): CAIP 2015, Part I, LNCS 9256, pp. 642–653, 2015.
DOI: 10.1007/978-3-319-23192-1_54

In the field of face tracking, a lot of studies over the last decade has provided significant progress, such as: i) using template models, e.g, Active Appearance Model (Cootes et al. [2001], Xiao et al. [2004], Matthews and Baker [2004]), Cylinder (Cascia et al. [2000], Xiao et al. [2003]) or Candide (Alonso et al. [2007]), ii) using local matching, e.g, (Vacchetti et al. [2004], Jang and Kanade [2008], Wang et al. [2012]), iii) using local discriminative classifiers, e.g, Constrained Local Model (Cristinacce and Cootes [2006], Wang et al. [2008], Saragih et al. [2011]). In recent years, cascaded regression has become the leading approach for accurate and robust face alignment, in which most of them has achieved state-of-the-art performance. The basic idea is to use a sequence of weak regressors, which are learned sequentially. The pioneer works of cascaded regression have been proposed in (Dollar et al. [2010], Valstar et al. [2010]) for object alignment, then be applied in (Cao et al. [2012]; Xiong and la Torre Frade [2013]) successfully for face alignment and tracking. Many methods were proposed using cascaded regression to improve performances in terms of accuracy, speed, or occlusion (Burgos-Artizzu et al. [2013], Ren et al. [2014], Kazemi and Sullivan [2014], Zhang et al. [2014], Sun et al. [2013]). Cascaded regression based methods have shown the high accuracy and real-time speed, merely it is restricted between (-45°,45°). This problem happens because of two reasons: First, the acquisition of ground-truth for unconstrained views is too expensive in practice. Second, annotating hidden landmarks at invisible side is difficult. In this paper, we will base on this approach and propose a solution to make it work at larger poses.

In the literature, some datasets were reported for pose estimation evaluation in video sequences. The most popular one is Boston University Face Tracking (BUFT) dataset (Cascia et al. [2000]). Its ground-truth of 3D pose is captured by magnetic sensors "*Flock and Birds*" with an accuracy of less than 1^o. It has two subsets: uniform-light set and varying-light set. The uniform-light set has a total of 45 video sequences (320×240 resolution) for 5 subjects (9 videos per subject) with available ground-truth of pose consisting of three directions: Yaw, Pitch and Roll. The varying-light set contains 27 sequences of 3 subjects recorded under same condition like the first set except fast-changing lighting conditions. So far, this is no longer a challenging dataset, because all subjects in this dataset always kept their faces neural while moving their heads slowly and no occlusion happens, the background is not too cluttered and the angles of three direction is mostly not larger than 40°. Many works reported high performance on this dataset (Xiao et al. [2003], Jang and Kanade [2008], Wang et al. [2012]). CLEAR07 (Stiefelhagen et al. [2007]) contains multi-view video recordings of a seminar room. It consists of 15 videos with four-synchronized cameras with frame rate at 15fps. This dataset provides both pose data from single view and multi-view. However, the subject captured for head pose is just seating in the same place. IDIAP Head Pose (Ba and Odobez [2007]) is an another source of LEAR07, and the values of Yaw, Pitch and Roll range only between (-60,60), (-60,15) and (-30,30) respectively for the single view. Methods evaluated on this dataset not fully automatic because the bounding-box is provided.

(Murphy-Chutorian and Trivedi [2008]) created a dataset recording drivers while driving in daytime and night-time lighting conditions. The drivers faces are usually neutral at mostly frontal views. There are other datasets reviewed in the survey (Murphy-Chutorian and Trivedi [2009]), but they are not much more challenging than above-mentioned datasets because of simple background, slow motion or frontal. Most of datasets reported used the magnetic sensors not comfortable to wear and move around because of wires connecting between them and the computer. In this work, we will propose a new way of recording 3D ground-truth that enables people to move around more comfortably. By this, we create a new database to evaluate in-the-wild pose tracking methods efficiently and more detail.

So, our contributions in this paper consists of two folds: i) We implement the state-of-the-art face tracking method and then propose a new way that allows to able to work with extreme poses or in-the-wild conditions. ii) We propose a new recording system to do recording accurately and more comfortable. Our new database contains challenging conditions that enable to evaluate advances of 3D pose tracking. The remaining of this paper is organized as follows: Section 2 describes the background of the face tracking method we want to improve. Section 3 discusses about our recording system. Experimental results and analysis are presented in Section 4. Finally, we provide in Section 5 some conclusions and further perspectives.

2 Towards to In-the-wild Tracking

Cascaded regression approach is now a promising approach which has shown the reliable capability of in-the-wild face tracking, but one of its limitations is out-of-plane tracking. In this study, we base on this approach and propose a new way to let it work at larger rotation.

2.1 Cascaded Regression

Let the shape $\mathbf{S} \in R^{2p\times 1}$ be the coordinate vector of p facial landmarks. Let $\widehat{\mathbf{S}}$ be true shape. The goal of face alignment is to align the shape $\mathbf{S}$ as closely as the true shape by minimizing $||\widehat{\mathbf{S}} - \mathbf{S}||$. To estimate the shape, we use a sequence of T weak regressors $(r(1), r(2), ..., r(T))$, $r(t) \in R^{2p\times D}$, D is the dimension of feature $\Phi \in R^{D\times 1}$ extracted from the image:

$$\mathbf{S}(t) = \mathbf{S}(t-1) + r(t)\Phi(I, \mathbf{S}(t-1)), \quad t = 1, ..., T \tag{1}$$

Given the facial image I and the initial face shape $\mathbf{S}(0)$, one weak regressor $r(t)$ estimates the new shape at time t $\mathbf{S}(t)$ from image features $\Phi(I, \mathbf{S}(t-1))$ computed using the previous shape $\mathbf{S}(t-1)$ on the image I. The sequentially training of the weak regressors $r(t)$ is based on N examples $\{(I_i, \widehat{\mathbf{S}}_i)\}_{i=1}^{N}$ by minimization as:

$$\arg\min_{r(t)} \sum_{i=1}^{N} \left\| \widehat{\mathbf{S}}_i - (\mathbf{S}_i(t-1) + r(t)\Phi(I_i, \mathbf{S}_i(t-1)) \right\|_2^2 \tag{2}$$

Φ is the vector concatenating local descriptors of p landmarks $\Phi = [\phi_1^T ... \phi_p^T]^T$. Hence, the local descriptor $\phi_i \in R^{d\times 1}$ is SIFT descriptor extracted in local region of i-th landmark. The initial shape $\mathbf{S}(0)$ is simply the mean shape aligned in the area detected by the face detectors. This minimization is a linear regression problem whose solution is in closed-form. We use the same way like (Xiong and la Torre Frade [2013]) to train the sequence of weak regressors. The parameters of our implementation, such as: patch size, the number of regressors,... is similar to (Xiong and la Torre Frade [2013]) and we obtained the same performance as in this paper.

2.2 Synthetic Data for Out-of-plane Tracking

Although cascaded regression approach is just considered for frontal face tracking so far, we believe that it is possible to work at out-of-plane rotations if a good training dataset of such conditions is provided. In the literature, two relevant datasets can be mentioned: AFLW (Koestinger et al. [2011]) and Multi-PIE (Gross et al. [2010]). The main limitation of these datasets are no information of hidden landmark annotation on large Yaw images. So, the number of landmarks among views is different that makes a gap of tracking and pose estimation, for example, from frontal to profile views. Indeed, the view-based models like using mixture of trees as (Zhu and Ramanan [2012]) could be a solution, otherwise pose detectors are required. It turns out that most of methods usually solve this profile problem by adaptive ways. The state-of-the-art, Intraface (Xiong and la Torre Frade [2013]), was only capable to handle restricted ranges, e.g, Yaw $\in$ [-45° 45°]. (Saragih et al. [2011]) developed a tracker tracking larger Yaw but not too robust. In addition, Pitch and Roll are not well-considered in current datasets.

In contrast, the synthetic data is a cheaper solution to create a high pose variation database. Thanks to 3D face models, the landmark number are the same among views because 2D projection of hidden landmarks could still be located. So, the gap of multi-view tracking could be bridged. In this study, we propose to use 3D face reconstruction to create extreme poses and combine with real data. We suppose that the real dataset is good for frontal tracking while the synthetic dataset is for other views. The extra mount of synthetic images will be included to the current datasets for training. We implemented 3D Morphable Model (3DMM) (Blanz and Vetter [1999]) to create our own synthetic data. We annotated once 51 inner landmarks similar to 300-W dataset (Sagonas et al. [2013]). The boundary landmarks are not considered because textures of two face sides are not good in our implementation of 3D face reconstruction. {Fig. 1} are some examples of 51 annotated landmarks and its rendering in different poses. In detail, we generate randomly about 3000 synthetic data using 3DMM for extreme poses. For real data, we use about 2000 images of training subsets of LFPW and Helen from 300-W dataset like (Ren et al. [2014]). In fact, this algorithm is originally developed for 2D face alignment, but it could be extended easily for tracking by using the aligned face at previous frame as the initial for the current one. The 3D pose is estimated by using POSIT (Dementhon and Davis [1995])

Fig. 1. The annotation of 51 landmarks in synthetic images and its rendering.

aligning a 3D model into 2D shape in this study. Indeed, the pose estimation would be improved more if using Bayesian tracking techniques, e.g. (Ababsa [2009]; Ababsa and Mallem [2006])

3 Recording System for Unconstrained 3D Pose Tracking

To build one dataset, the most importance is the ground-truth. We propose to use one stereo-infrared system to capture the 3D pose. The system enables to capture the ground-truth accurately while the face pose is unconstrained. For the recording campaign, three things need to be well-prepared: the setup of recording system, the calibration of cameras, and the definition of recording protocols.

Recording System. The proposed system consists of one RGB camera and one stereo-infrared device (SMARTTRACK) installed as {Fig. 2}. SMARTTRACK could detect five markers of the fly-stick from a distance and estimate accurately its 6 DOF. To use the fly-stick for the head pose ground-truth, we design a fly-stick hat people can wear on as {Fig. 2} while moving around. The system is installed as follows: The infrared device is located about 50cm higher than the RGB camera to be able to always detect all markers. The tripod enables to change the system height respect to people being recorded. These cameras are kept fixed vertically during the recording process and connected to one computer via the wire cable. For the infrared device, the driver is available on the website of the provider SMARTTRACK and it provides also the C/C++ API and the software Dtrack2 to connect and control. Notice that the infrared device and RGB camera need to be synchronized (we developed a function for this purpose) to capture at the same frame rate. The RGB camera and the infrared system also needs to be calibrated to get the correct 3D ground-truth in the coordinate of RGB camera. The calibration process of two devices is presented in following section.

Calibration. The calibration is first performed for each device. The traditional calibration using the check-board 7×9 via the Matlab toolbox (Bouguet [2003]) for RGB camera. The Dtrack2 supports the automatic calibration process for infrared device. For stereo calibration, we do annotation the 2D location of

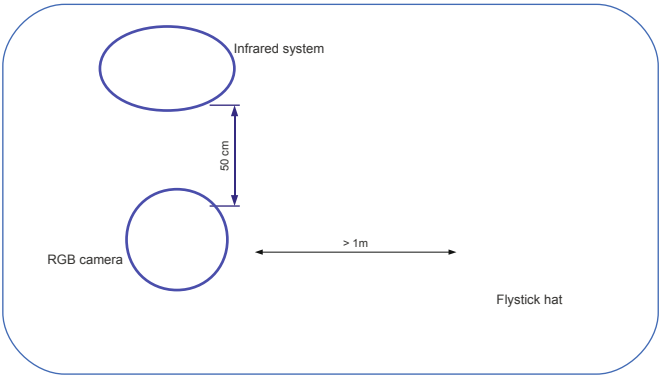

Fig. 2. The installation of recording system.

reference origin (red point in {Fig. 3}) on 2D images responding to 3D location detected by infrared cameras. So, 45 frames are collected to annotate the origin of the fly-sticks with available 3D ground-truth of infrared device. The {Fig. 3} shows the stereo-calibration to estimate the rotation $\mathbf{R}_{ic}$ and translation $\mathbf{T}_{ic}$ from the infrared to RGB cameras. Let denote that the i-th frame has 2D position of origin is $\mathbf{l}_i$ corresponding to the its known 3D coordinates $\mathbf{L}_i$ estimated by the infrared device. The transformation matrices are estimated through the least square problem:

$$\{\hat{\mathbf{R}}_{ic}, \hat{\mathbf{T}}_{ic}\} = \arg\min_{\mathbf{R}_{ic}, \mathbf{T}_{ic}} \sum_{i=1}^{N} \left(\mathbf{l}^i - \mathscr{P}([\hat{\mathbf{R}}_{ic}\hat{\mathbf{T}}_{ic}]\mathbf{L}^i)\right)^2 \tag{3}$$

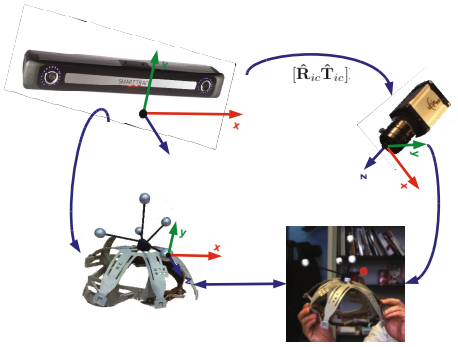

Fig. 3. The calibration diagram of RGB and infrared cameras.

where $\mathscr{P}$ is the projection of 3D position of fly-stick after transforming it from the infrared device to the RGB camera's coordinates. However, the location

and rotation estimated is only the movement of the fly-stick worn on the head. It is difficult to have exactly the rigid motion of the head because the head sizes of people are different. To solve this problem, we define the pose of the first frame as the reference of (Yaw,Pitch,Roll)=(0,0,0). It means the first frame needs to make sure always to be frontal. The coming frames computed as the difference from the first one. By this way, the movement of the head is exactly the movement of the fly-stick. The calibration error of RGB camera using Matlab calibration toolbox is (0.162, 0.161) pixels in horizontal and vertical directions, and the stereo-calibration error between RGB camera and infrared system that we proposed is 3 pixels. It is acceptable value that is not too critical because we use the first frame as a reference for to compute ground-truth of next ones.

Recording. Our dataset consists of 50 videos of 10 subjects. One subject is recorded five videos during 20-40 seconds/video with the resolution 768×576. The first frame is always the frontal respect to the camera view. We set up the recording protocol to capture videos in challenging conditions: illumination, wide rotation, expression, occlusion, fast motion and complex movements. The dataset is recorded in the office environment with the cluttered background. Videos are recorded as the following protocols in raising up the difficult levels gradually for each subject from video 1 to 5:

- The 1st video (frontal): Subjects stay in front of the camera system and then walk around in the view of camera while keeping looking at it.
- The 2nd video (near-frontal + expression): Like the first one; except subjects sometimes change their expression or rotate a little bit head directions.
- The 3rd video (profile): Subjects stay in front of the camera system, then walk around and change its head direction, possibly wide rotations.
- The 4th video (profile + expression + occlusion): Like the third one, except subjects sometimes change expression or make some occlusions.
- The 5th video (in-the-wild): Subjects move comfortably in the view of camera.

As result, one video has its ground-truth file. The format of each line in ground-truth consists of 6 values [V_x V_y V_z Yaw Pitch Roll]. V_x, V_y, V_z: 3D coordinates of three axis of x, y, z and Yaw, Pitch and Roll are angles of three orientations respectively. Notice that V_x, V_y, V_z and Yaw Pitch Roll is computed corresponding to the reference of the first frame. At some frames, the infrared camera could not find out the fly-stick, so all the field of rotation and translation are set to -1 at these frames. For evaluation, we propose two measures: robustness and accuracy. The robustness P is the percentage of the number of well-detected frames N_s over the total of frames of ground-truth N_{total} (excluding frames having no ground-truth): $P = \frac{N_s}{N_{total}} 100\%$. One frame is well-detected if the distance of the 2D position (projection) between our estimation and ground-truth is smaller than a threshold, and we fix this threshold be 40 pixels in our experiments. Notice that to be robust with scale problem, we normalize the distance by the factor $\frac{V_z}{F_{focal}}$, where F_{local} is the focal length of RGB camera.

Fig. 4. Some sample video sequences in our datasets.

The accuracy is the Mean Absolute Error (MAE) between the estimation and the ground-truth of head pose. For example, the error of Yaw (similarly for the Pitch and Roll) are computed as follows:

$$E_{yaw} = \frac{1}{N_s} \sum_{i \in S_s} |\Theta^i_{yaw} - \widehat{\Theta}^i_{yaw}| \tag{4}$$

where N_s is the number of well-detected frames and S_s is the set of frames Θ^i_{yaw} and $\widehat{\Theta}^i_{yaw}$ is the estimation and the ground-truth of Yaw. The error of the dataset is the average of all videos. {Fig. 5} shows sample frames of our database.

4 Experiments

4.1 Public Datasets

Before testing methods on own recordings, we report their performance on uniform-light set of BUFT. Our purpose is to make sure that state-of-the-art methods chosen to investigate our recordings later are good enough. In {Table 1}, although we are not better than some methods reported on this dataset, we have the more or less the same performance; especially, the Yaw and Pitch precision compared to fully-automatic method ("*" is fully automatic, otherwise semi-automatic). Among methods that can estimate simultaneously rigid and non-rigid parameters ("+" is able to estimate the non-rigid parameters), we are much better than FaceTracker (Saragih et al. [2011]). It demonstrates the efficiency of our tracking approach. Intraface (Xiong and la Torre Frade [2013]) is the best among fully-automatic methods. In fact, it is similar to ours except it is trained on much more real data than ours (but no detail is reported). Notice that the results of FaceTracker and Intraface on BUFT are performed by us.

To verify the in-the-wild and extreme pose tracking of our approach, two datasets are chosen: a) YouTube Celebrities dataset (Kim et al. [2008]), a challenging dataset as for in-the-wild conditions, b) Honda/UCSD (Lee et al. [2003]) with the complex motion patterns and extreme poses. This video[1] shows the results on some videos that our method is robust with these challenging videos.

[1] http://goo.gl/PL11JC

Table 1. The tracking performance on the uniform-light set of BUFT dataset.

Approach	E_{yaw}	E_{pitch}	E_{roll}	E_m
Wang et al., 2012	3.8	2.7	1.9	2.8
Xiao et al., 2003	3.8	3.2	1.4	2.8
Intraface (*,+)	4.1	3.0	2.2	3.1
Lefevre and Odobez, 2009 (+)	4.4	3.3	2.0	3.2
Jang and Kanade, 2008 (*)	4.6	3.7	2.1	3.5
Asteriadis et al., 2014 (*)	4.3	3.8	2.6	3.5
Morency et al., 2008 (*)	5.0	3.7	2.9	3.9
Facetracker (*,+)	4.3	4.8	2.6	3.9
Our method (*,+)	**4.4**	**3.3**	**3.0**	**3.6**

4.2 Unconstrained 3D Pose Tracking (U3PT) Dataset

In this section, we will use our own dataset (U3PT) to evaluate our tracker and two other trackers: FaceTracker and Intraface, two fully-automatic methods, that have good results on BUFT. Intraface based on the frontal face detectors to align the face, so they are only good for near-frontal views. FaceTracker can track the profile but be not really robust. Otherwise, our method can track well the high number of frames compared to other methods. The results in {Table 2} show performances of considered methods. We have the best results for the robustness but we are worse than Intraface for precision.

Table 2. The performance of methods on our own U3PT dataset.

Approach	P	E_{yaw}	E_{pitch}	E_{roll}	E_m
FaceTracker	30.6%	9.9	10.3	8.9	9.7
IntraFace	52.1%	6.2	6.4	7.0	6.5
Our method	**54.2%**	**8.4**	**7.8**	**6.3**	**7.5**

To be more detail, we evaluate considered methods on each specific groups of our database. Each group has a similar head movement patterns as reported in the earlier section. It enables us to see more detail the capability of methods on challenging cases. {Fig. 5} shows the accuracy and robustness of methods on each group. FaceTracker is the worst in the comparison at both robustness and precision in all groups. Two first groups, we have best performances on both accuracy and robustness. It means that our methods are more robust with frontal and expression tracking compared to other methods. However, we have only better robustness at three remaining groups compared to Intraface. It is because three groups have profile rotation that seems our estimation of profile views are not really good. Although our method has a better robustness but lower accuracy compared to Intraface.

The performance let us know about the difficulty of our recordings. Only about a half of number frames can be well-tracked, while the accuracy is low. Our method could work well with out-of-plane rotation on UCSD/Honda dataset, but it is impossible to do the same on U3PT because likely the background of this dataset is too cluttered. Whereas, our synthetic training data is completely black.

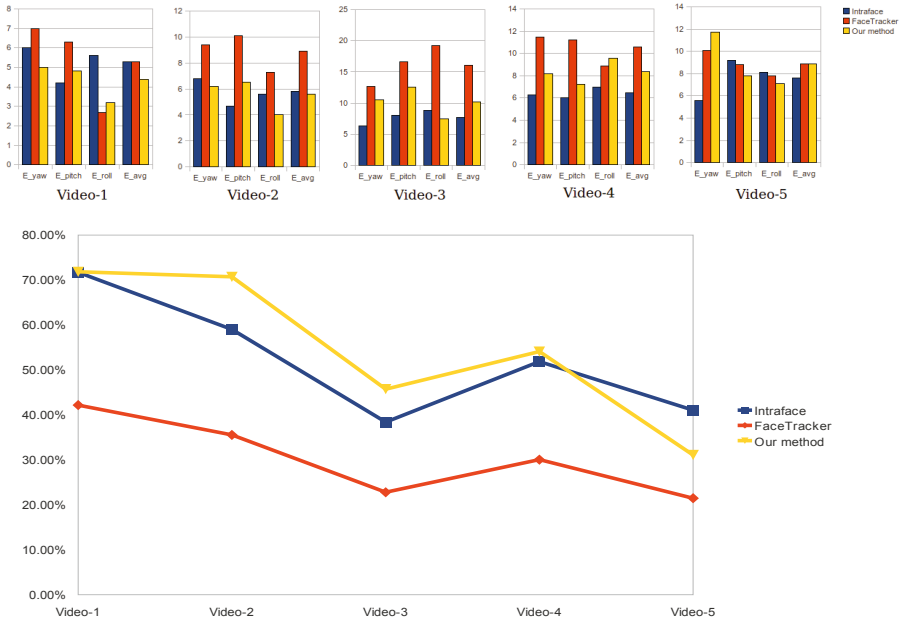

Fig. 5. The accuracies and robustness on each video group.

Intraface detects only a half of number frames because it bases on the frontal face detection to align 3D face. Although, the number of our real training data is only about 2000 images too small compared to Intraface, but with the support of synthetic data could make our method comparable to it. We believe if 3DMM is implemented better (background and expression), our method will be much better.

5 Conclusion

We proposed a new approach using cascaded regression for efficient tracking using the combination of synthetic and real dataset. The real dataset is good for frontal view, otherwise, the synthetic data makes the system robust with the large of rotations such as: Yaw or Pitch. The result shows the possibility of using synthetic data to by-pass the difficulties of lack of training data for such situations as Honda/UCSD video demo. By the combination of two kind of datasets, the robustness with large rotation tracking is more robust. However, when using on cluttered background, our synthetic images is not useful because its background is completely black. In addition, the facial animation of our 3DMM model haven't yet implemented to make the face tracking more reliable. In the future work, we aims to develop 3DMM with facial expression

and background to create better databases. Furthermore, we also propose the campaign of ground-truth of 3D pose for in-the-wild tracking dataset allows the evaluation of the accuracy and robustness of tracking methods with challenging conditions. It would be useful as a benchmark for tracking methods in the future.

References

Ababsa, F.: Robust extended kalman filtering for camera pose tracking using 2d to 3d lines correspondences. In: IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 1834–1838 (2009)

Ababsa, F., Mallem, M.: Robust line tracking using a particle filter for camera pose estimation. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology (2006)

Alonso, J., Davoine, F., Charbit, M.: A linear estimation method for 3d pose and facial animation tracking. In: CVPR (2007)

Asteriadis, S., Karpouzis, K., Kollias, S.: Visual focus of attention in non-calibrated environments using gaze estimation. IJCV (2014)

Ba, S.O., Odobez, J.-M.: Probabilistic head pose tracking evaluation in single and multiple camera setups. In: Classification of Events, Activities and Relationship Evaluation and Workshop (2007)

Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH, pp. 187–194, New York, NY, USA (1999)

Bouguet, J.Y.: Camera calibration toolbox for matlab (2003)

Burgos-Artizzu, X., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: ICCV (2013)

Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR (2012)

Cascia, M.L., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. TPAMI **22**(4), 322–336 (2000)

Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. TPAMI **23**(6), 681–685 (2001)

Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: BMVC (2006)

Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. IJCV **15**, 123–141 (1995)

Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression. In: CVPR (2010)

Gross, R., Matthews, I., Cohn, J.F., Kanade, T., Baker, S.: Multi-pie. IVC **28**(5), 807–813 (2010)

Jang, J.-S., Kanade, T.: Robust 3d head tracking by online feature registration. In: FG (2008)

Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR (2014)

Kim, M., Kumar, S., Pavlovic, V., Rowley, H.A.: Face tracking and recognition with visual constraints in real-world videos. In: CVPR (2008)

Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (2011)

Lee, K., Ho, J., Yang, M., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds **1**, 313–320 (2003)

Lefevre, S., Odobez, J.-M.: Structure and appearance features for robust 3d facial actions tracking. In: ICME (2009)

Valstar, M.F., Martinez, X.B., Pantic, M.: Facial point detection using boosted regression and graph models. In: CVPR, pp. 2729–2736 (2010)

Matthews, I., Baker, S.: Active appearance models revisited. IJCV **60**(2), 135–164 (2004)

Morency, L.-P., Whitehill, J., Movellan, J.R.: Generalized adaptive view-based appearance model: integrated framework for monocular head pose estimation. In: FG (2008)

Murphy-Chutorian, E., Trivedi, M.M.: HyHOPE: Hybrid Head Orientation and Position Estimation for Vision-based Driver Head Tracking. IEEE Intelligent Vehicles Symposium (2008)

Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. PAMI **31**(4) (2009)

Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features (2014)

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: ICCV Workshops (2013)

Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. IJCV **91**, 200–215 (2011)

Stiefelhagen, R., Bernardin, K., Bowers, R., Rose, R.T., Michel, M., Garofolo, J.S.: The CLEAR 2007 evaluation. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) RT 2007 and CLEAR 2007. LNCS, vol. 4625, pp. 3–34. Springer, Heidelberg (2008)

Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR (2013)

Vacchetti, L., Lepetit, V., Fua, P.: Stable real-time 3d tracking using online and offline information. TPAMI **26**(10), 1385–1391 (2004)

Wang, H., Davoine, F., Lepetit, V., Chaillou, C., Pan, C.: 3-d head tracking via invariant keypoint learning. IEEE Transactions on Circuits and Systems for Video Technology **22**(8), 1113–1126 (2012)

Wang, Y., Lucey, S., Cohn, J.: Enforcing convexity for improved alignment with constrained local models. In: CVPR (2008)

Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR (2012)

Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2d+3d active appearance models. CVPR **2**, 535–542 (2004)

Xiao, J., Moriyama, T., Kanade, T., Cohn, J.: Robust full-motion recovery of head by dynamic templates and re-registration techniques. International Journal of Imaging Systems and Technology **13**, 85–94 (2003)

Xiong, X., la Torre Frade, F.D.: Supervised descent method and its applications to face alignment. In: CVPR (2013)

Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-Fine Auto-Encoder Networks (CFAN) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 1–16. Springer, Heidelberg (2014)