2D Hand Tracking with Motion Information, Skin Color Classification and Aggregated Channel Features

J. H. Hammer[1,2], C. Qu[1,2], M. Voit[2], and J. Beyerer[2,1]
[1]Vision and Fusion Laboratory, Karlsruhe Institute of Technology Karlsruhe, Germany
[2]Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany

Abstract— *In this paper we present our latest approach for 2D hand tracking in video streams of head-worn monocular color cameras. Those are lightweight and embedded in many head-mounted devices (HMDs) like eye trackers, Augmented Reality glasses or Virtual Reality devices to capture the field of view from the ego perspective. Interaction with virtual elements of the augmented or virtual reality or the interaction with the device itself can be intuitively performed using the hands. Our new approach called AfM fuses our previous method called Motion Segmentation and Appearance Change Detection based Skin color detection (MACS) with the results of a hand detector using Aggregated Channel Features. It tracks the hand more robustly and reduces the deviation from the ground truth paths by 50 % on our benchmark with more than 25,000 frames consisting of different hand gestures.*

Keywords: hand tracking, hand gestures, object detection, skin color, optical flow

1. Introduction

Virtual reality (VR), as it is realizable already today, is going to change how we will be entertained, how we will learn and how we will collaborate. One day Augmented Reality (AR) devices will have overcome their current limitations of weight, small field of view and unimpressive visualization capabilities. Those devices will replace all sorts of auxiliary monitors as tablets or smartphones which are currently often used e.g. in industry or museums to bring information closer to the user. The user then has to capture this information and transfer it to the real object. With AR the step of transferring the information is not needed any more, since information can be visualized referenced to the world directly in the field of view on the objects. AR glasses have already proven its necessity in different applications like e.g. the picking process in warehouses[19] or industrial assembly since the hands are free when no device needs to be hold in the hands.

When the visualization capabilities of AR glasses will be impressive enough in the near future one wants to interact with the virtual contents visualized in front of oneself. Hand movements and gestures are the most intuitive modality for manipulating such virtual contents. In today's VR applications controllers are used instead of the hands. Having no need for those controllers would be a great benefit. Therefore, the hands must be tracked reliably even in difficult environments with highly challenging lighting conditions. Many researcher focus on hand pose estimation using depth data [1], [2], [21], [22], but in both, industrial and commercial applications, the weight of the HMD is one of the most important factors to allow for daily usage in real life applications. Stereo cameras and depth sensors always add additional weight and especially construction complexity. That is why many VR devices have only one monocular camera embedded. Hence, monocular hand tracking remains important to be focused on [3], [5], [6], [7], [20].

In the following sections we will first concentrate on related work (Section 2). Then we will summarize one of our previous 2D hand tracking approaches that yields as basis of the work on hand in Section 3.1. Afterwards we present the created hand detector in Section 3.2 and show in Section 3.3 how we embedded it into our previous approach to achieve a much more accurate hand tracking, which is quantified in the evaluation in Section 4. We end with a conclusion in Section 5.

2. Related Work

In mobile applications with a head-mounted camera, the background is not static. Hand localization can therefore not be achieved by simple frame subtraction. The lighting conditions may change all the time and direct sunlight can be illuminating the scene. 3D sensors are widely and successfully used for hand tracking and hand pose estimation [1], [2] because depth information easily results in an accurate segmentation of the scene. Nowadays active depth sensors are not heavy any more but are restricted to indoor applications. A light-weight passive stereo camera system would be another option to acquire depth information, but robust stereo reconstruction heavily increases the computational load. To sum up, every type of depth sensor adds additional needs to the size and weight of a HMD as well as complexity concerning the electrical wiring and increases power consumption. Accordingly, it is achievable to use only one single RGB camera. Gloves [9], markers [10], accelerators [11] or thermal cameras [12] have been used. But the optimal solution would be to not need to attach further devices to the hands of the user. Pisharady et al. [6] proposed a method for hand posture detection

from single view even in front of a complex background but their method is not applicable for real-time processing. Hammer and Beyerer [7] compared different hand tracking methods with already established approaches [13][14]. In [8] we presented a 2D hand tracking approach called Motion Segmentation and Appearance Change Detection based Skin color detection (MACS) where we embedded the pixelwise skin-color classification approach of Li et al. [5] to better handle different environments with different lighting conditions. Since our new approach uses MACS as basis we will summarize the most important steps of MACS in a part of the next section. For a detailed explanation of MACS we refer to [8]. The main contribution of this work is the training of a hand detector based on Aggregated Channel Features for object detection presented by Dollar et al.[25] and the fusion process of its hand position candidates with the hand position computed by MACS, which decreases the tracking deviation by about 50% on our benchmark also presented in [8].

3. Hand Tracking Using Motion, Color and Aggregated Channel Features

The new hand tracking approach called AfM (ACF feat. MACS) relies on MACS and refines its hand position guess with the candidates computed by the hand detector using Aggregated Channel Features (ACF). In the following sections we briefly summarize MACS, describe the hand detector and the tested developed strategies.

3.1 Brief Summary of MACS

For a detailed description of MACS we refer to [8]. Its workflow can be summarized as follows: First the optical flow is estimated (Figures 1a, 1b, 1c) and clustered. From the determined clusters the foreground region assumed to contain the hand is extracted (view Figures 1d, 1e) using k-means and knowledge about the size of the hand in the images of a head-worn camera. To more robustly track this motion segment, we observe its appearance computed from the masked image (Figure 1f). This technique is essential for working with corrupted motion information especially at image boarders and when no optical flow is estimated or could not be clustered. We named this *Appearance Change Detection*. The skin color classification approach of Li et al.[5] is a further indicator for the hand segment. MACS fuses the resulting skin color segmentation (Figure 1g) and the moving foreground segmentation yielding a fused segmentation (Figure 1h) by only keeping those blobs of the skin color segmentation that overlap with areas of foreground motion. This fused segmentation is the basis for the tracking with the *shape*-particle filter (Figure 1i) that was introduced in [7]. The hand position is determined as the median position of all particles. If the average amount of skin-colored pixels per particle in a square local neighborhood of 50×50 pixels centered at the particle's position is above a certain threshold (750 pixels in our case), the hand is assumed to be in the image. Then enough skin colored pixels are in place of the particles. Otherwise the hand is assumed as not visible. This threshold is one of MACS limitations since a wrong segmentation can easily produce false positives.

3.2 Hand Classification Using Aggregated Channel Features

In order to facilitate hand detection the Aggregated Channel Feature (ACF) algorithm [25] is employed, which exploits rich color feature descriptors in conjunction with an effective classifier, following the classical approach of the Histograms of Oriented Gradients (HOG) [26].

Specifically, rather than solely relying on the gray-scale information, ACF leverages the LUV color channels combined with the gradient magnitude and 6 HOGs, leading to totally 10 feature channels. Pre- and post-smoothing with Gaussian filtering followed by 4×4 pooling are conducted to aggregate the pixels for compact and efficient representation.

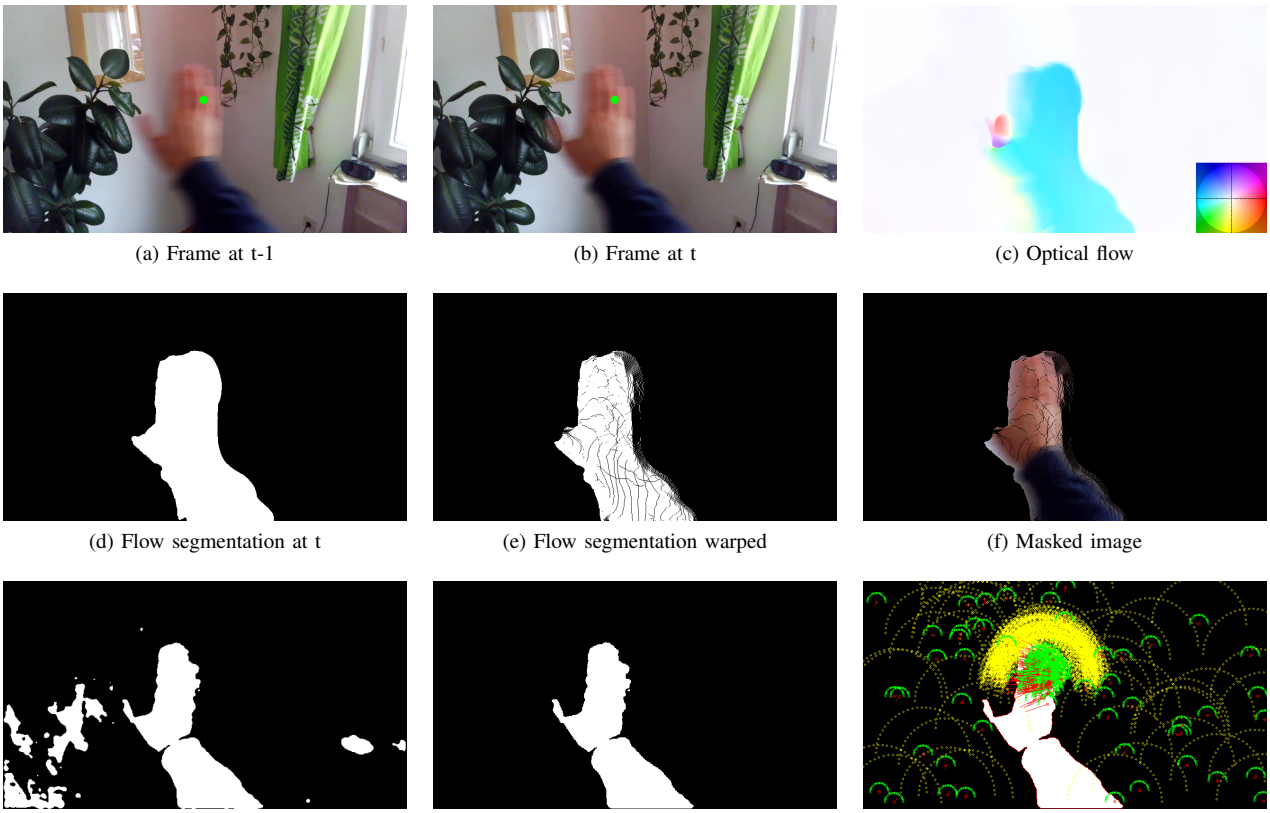
To circumvent the computational complexity of recomputing dozens of different feature pyramids, ACF only performs few of them with exact values and approximates both the color and HOG channels for the remaining scale spaces by simple rescaling of the known ones, which is justified by the power law of the natural image statistics.

Training is carried out on hand crops of 50×50 pixels extracted from a subset of frames of our sequences. Since the problem setup resembles that of the original pedestrian detection task with the palm like the upper body and the non-rigid fingers analog to the legs, we apply the same 2048 depth-two boost trees for learning with bootstrapping.

Example output of the hand detector can be seen in Figure 2. The hand detector produces results on the hands and fingers but unfortunately also on the arm and the background. Simply using the candidate with the highest score is therefore no option. But we can use MACS as guess for the hand location and refine it by the ACF candidates as described in the next section.

3.3 Fusion of MACS and the ACF-Based Hand Detector

To fuse MACS and the ACF-based hand detector we need to look at what both produce: On the one side there is the hand position guess $h_{MACS} \in \mathbb{R}^2$ produced by MACS. Furthermore MACS can decide that no hand was found. On the other side there are many hand candidates i centered at pixel p_i computed by the ACF-based hand detector where each guess comes with an associated score s_i. And as we have seen there are often guesses computed even if no hand is visible. Fortunately those guesses come with low scores.



(a) Frame at t-1 (b) Frame at t (c) Optical flow

(d) Flow segmentation at t (e) Flow segmentation warped (f) Masked image

(g) Color segmentation (h) Fused segmentation (i) Tracking with *shape*-particle filter

Fig. 1
MACS OVERVIEW

Below we are describing different strategies for the fusion process.

3.3.1 Simple MACS Refinement

This method is the most simple approach for using the hand detector guesses. The MACS guess h_{MACS} is refined by a hand detector guess i_{best} that is close to h_{MACS} and at the same time has a high score. This i_{best} is computed as

$$i_{best} = \arg\max_i c_i \tag{1}$$

with $i \in \{1, \ldots, n\}$ based on a confidence score c_i with

$$c_i = d_i + v_i \tag{2}$$

consisting of a distance weight d_i

$$d_i = 1 - \frac{||p_i - h_{MACS}||_2}{\sum_j ||p_j - h_{MACS}||_2} \tag{3}$$

with $j \in \{1, \ldots, n\}$ and a score weight v_i

$$v_i = \frac{s_i}{\sum_j s_j}. \tag{4}$$

The final hand position h_t at the current point of time t is then set to be $p_{i_{best}}$. In the presented work we considered at the maximum the ten best hand detector guesses per frame ($n = 10$). If no guess is computed, the hand is assumed to have left the image.

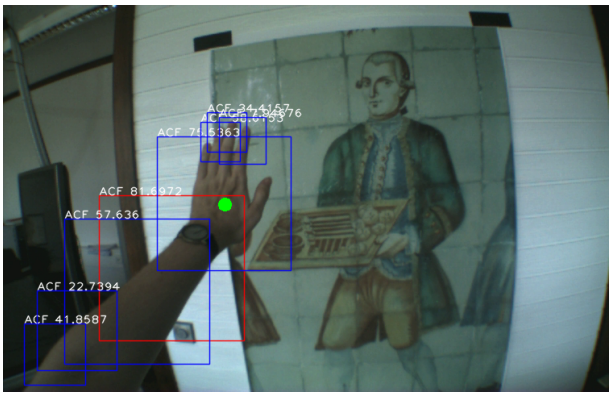
3.3.2 Refinement with Propagation

Since the MACS guess h_{MACS} itself is often not accurate with a median distance to the ground truth path of 34 pixels [8] the distance weight (3) is weighting wrong hand detector guesses p_i better because they are closer to the MACS guess. To solve this problem we use a simple propagation step that takes the last two hand positions h_{t-1} and h_{t-2} into account. The distance weight (3) is then adjusted as

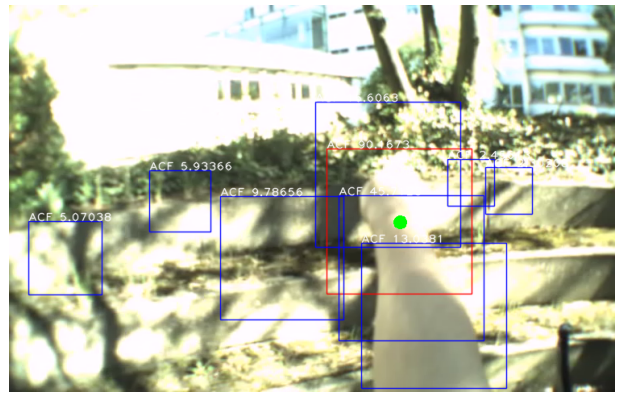
$$d_i = 1 - \frac{||p_i - h_g||_2}{\sum_j ||p_j - h_g||_2} \tag{5}$$

with

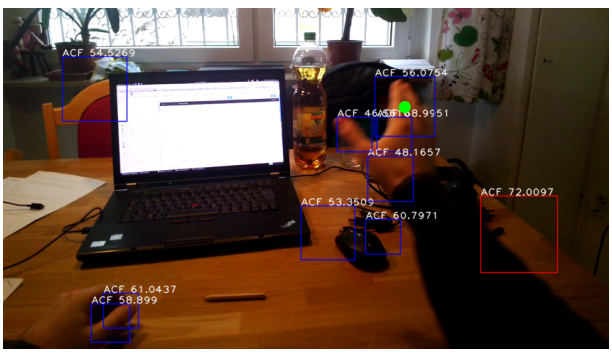
$$h_g = \begin{cases} h_{t-1} + (h_{t-1} - h_{t-2}) & \text{if } h_{t-1} \text{ and } h_{t-2} \text{ available} \\ h_{\text{MACS}} & \text{otherwise} \end{cases} \tag{6}$$



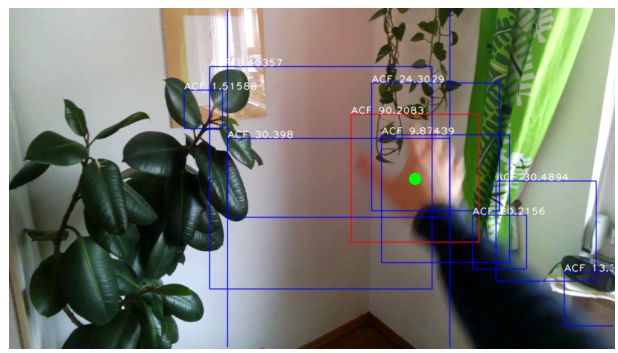
(a) Candidate with highest score of 81 is placed on arm but the second best with a score of 75 is at correct hand position. Other candidates with scores below 58 are distributed on fingers and arm.



(b) Image with overexposed areas. The candidate with the highest score of 90 is the correct hand locations. Other candidates are placed on arm and background with scores below 50.



(c) Correct hand position not under the ten best candidates. But at least some candidates are on the hand and their scores are above 45. The best candidate at the right table margin has a score of 72. The candidates close to the computer mouse have scores between 53 and 60.



(d) Hands are even detected at the wall but the correct hand location has the best score with 90 compared to 30 of the second best candidate.

Fig. 2

RESULTS OF THE HAND DETECTOR. AT MOST TEN CLASSIFICATION RESULTS ARE SHOWN. THE ONE WITH THE RED COLOR HAS THE HIGHEST SCORE.

and the rest of the computation equals the computation described in 3.3.1.

3.3.3 Robust Distance Weighting

The problem with the above described distance weight is, that due to the normalization the value d_i depends on the number of hand detector guesses p_i taken into account and their distances. Therefore we change the distance rating computation to be related to a maximally accepted distance $d_{\max}$. The adapted computation of d_i is as follows:

$$d_i = \begin{cases} 0 & \text{if } ||p_i - h_g||_2 \geq d_{\max} \\ 1 - \frac{||p_i - h_g||_2}{d_{\max}} & \text{otherwise.} \end{cases} \quad (7)$$

Accordingly, if the distances of a hand detector guess to the position h_g is greater or equal $d_{\max}$ the guess is penalized with a distance weight of 0. If it is 0, it gets the highest distance weight of 1 and if it is in between, the weight is linearly lowered. This makes the distance weight independent of the number and local distribution of the hand detector guesses. Our current implementation uses a $d_{\max}$ of 100 pixels, which is a rough guess of the maximal displacement of the hand position between two consecutive frames.

3.3.4 Median Rejection of Hand Detector Guesses

To add more robustness against hand detection guesses that are too far away from the current track, we can do even more than the robust distance weighting described previously. Additionally we compute the median of the last m hand positions and state that the next hand position must not have a higher distance to this median position than $d_{\max}$. Therefore, all hand detector guesses p_i with a higher distance to the median position are discarded. We consider $m = 3$ in our implementation to not be thrown back from the current state too much.

Table 1
RESULTS FOR THE COMPLETE BENCHMARK

	TPR	FPR	PREC	F1	ACC	qu25	qu50	qu75
MACS	0.907	0.015	0.980	0.942	0.950	19.7 px	34.1 px	52.4 px
AfM REF	0.908	0.007	0.990	0.947	0.955	8.1 px	13.6 px	22.8 px
AfM PROP	0.904	0.007	0.990	0.945	0.953	8.1 px	13.6 px	22.5 px
AfM PROP ROB	0.905	0.007	0.990	0.946	0.954	8.5 px	14.1 px	24.7 px
AfM ALL	0.911	0.005	0.993	0.950	0.958	8.1 px	13.5 px	22.5 px

3.3.5 Adaption of the Segmentation Mask

Since the particle filter used for producing the MACS hand position guess p_{MACS} gets a binary segmentation, we can simply adjust this binary segmentation by setting all pixels to background whose distance to the median position introduced in 3.3.4 is greater than some distance. Due to the shape particle we experienced good results with such a distance of the 1.5-fold of d_{max}, so in our case 150 pixels.

3.3.6 Rejecting Candidates based on the ACF Score

One of MACS limitations is that it is not tracking hands but a moving skin-colored foreground object. The decision if a hand is found is simply made upon the average number of skin pixels $\bar{s}_{\text{skin}}$ in the local 50×50 pixels neighborhood of each particle. If $\bar{s}_{\text{skin}}$ is above a certain threshold MACS assumes to have found a hand because this is the case when most of the particles condense on the hand blob. Unfortunately, when the segmentation is not good, this decision rule is not robust and highly adapted to the used data sets of [8]. We can improve this by instead using the ACF scores of the hand detector results. So, instead of using the threshold for the decision making if a hand is visible in the image or not, we set it to a very low value of $\bar{s}_{\text{skin}} = 100$ to let MACS find even more often a hand position guess. This produces more false positives but we can reject all hand detector guesses with lower scores than a threshold $t_{\text{ACF}} = 20$ to detect if a hand is visible or not.

4. Evaluation

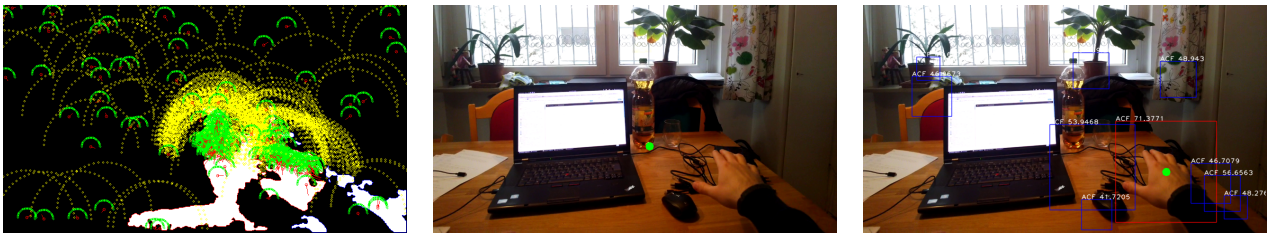
In [8] we presented our new benchmark consisting of 29 videos with more than 25,000 frames, different and challenging lighting conditions and wooden elements that make skin color classification difficult. We described our evaluation methodology and a good tracking result: The true positive rate (TPR) should be at least above 80 %. The false positive rate (FPR) should be as low as possible. The F1-measure (F1), the precision (PREC) and the accuracy (ACC) should be as high as possible. Additionally, the distances between all true positives and their corresponding ground truth hand positions should be as low as possible. But due to the subjective trajectory annotation [8] a deviation from the ground truth of less or about seven pixels per frame can be seen as perfect result. Even tracks with a deviation of up to 18 pixels per frame would subjectively be considered as very good tracking. To compare the accuracy of the tracking we are going to look at the 0.25 (qu25), 0.50 (qu50) and 0.75 (qu75) quantile of the distances.

In the evaluation we take into account from [8] only MACS since all other algorithms it was compared to failed to produce good tracking results due to their static skin color models. Then we compare MACS to the different variants of the new AfM-algorithm:

- MACS: Motion segmentation and Appearance Change Detection based skin color detection ([8])
- AfM REF: Simple MACS Refinement (Section 3.3.1)
- AfM PROP: Refinement with propagation (Section 3.3.2)
- AfM PROP+ROB: Robust Distance Weighting (Section 3.3.3)
- AfM ALL: Refinement with propagation (Section 3.3.2), robust distance weighting (3.3.3), median rejector (Section 3.3.4), mask adaption (3.3.5), ACF score rejection (Sections 3.3.6)

The detection rates, including the true positive rate, the false positive rate, the precision, the F1-measure, the accuracy, and additionally the statistics for the deviation from the ground truth path are computed in a manner as if all sequences would have been concatenated to one long sequence. The results are illustrated in Table 1.

MACS reaches a good true positive rate of 90.7 %. It shows a low false positive rate of 1.5 % and a high precision of 98 %. The F1-measure of 94.2 % is also high and the accuracy of 95.0 % is high as well. When looking at the AfM REF results, we see a better true positive rate of 90.8 %, which results in seven correctly detected hands more. The false positive rate of 0.7 % states that about 100 false positives less as MACS were produced. Accordingly the precision and the F1-measure are slightly better. The results of AfM PROP and AfM PROP ROB show that the refinement with propagation and the robust distance weighting on their own do not improve the results. Variant AfM ALL, using additionally the median rejection of hand detector guesses, the adaption of the segmentation mask and the rejection of ACF candidates based on their score, produces about 40 true positives more and 30 false positives less than AfM REF, resulting in an even better true positive rate of 91.1 %, a



(a) Condensation of the particles

(b) Input image with false hand position estimation by MACS

(c) Input image with correct hand position estimation by AfM ALL

Fig. 3

MACS FAILS BUT AFM TRACKS CORRECTLY

better false positive rate of 0.5 %, a precision of 99.3 % and an F1-measure of 95.0 %.

When we look at the distance values, we see the real difference between MACS and AfM All. The deviations from the ground truth of the tracks computed by MACS show a 0.25 quantile of 19.7 pixels, a 0.50 quantile of 34.1 pixels and a 0.75 quantile of 52.4 pixels. AfM ALL shows more than 50 % smaller deviations with a 0.25 quantile of 8.1 pixels, a 0.50 quantile of 13.5 pixels and a 0.75 quantile of 22.5 pixels. This tells us that AfM ALL produces a much more accurate tracking. Regarding that 18 pixels deviation are subjectively seen as accurate tracking [8], we can say that with AfM ALL we have come much closer to the goal of accurate tracking on our videos.

Comparing the variants of AfM the differences are marginal when looking at the complete benchmark. But the results also show that AfM ALL produces even better results without the need for a highly adapted threshold for the average number of skin pixels used by the particle filter for deciding if a hand is present or not. This decision has been delayed to the rejection of ACF candidates based on their score (view Section 3.3.6), and thus makes AfM ALL more general than the other AfM variants.

To show how AfM improves tracking accuracy we look at the example where MACS suffered from poor distinction of skin and wood as shown in Figure 3. The particles move to the upper left part of the segmentation result and the estimated hand position is distracted as visualized by the green dot in Figure 3b. This problem is solved by AfM because of its hand detector as can be seen in Figure 3c where the green dot is placed on the hand.

5. Conclusion

To sum up, we have shown a new 2D hand tracking algorithm called AfM that improves our previous method MACS by the usage of hand detector based on Aggregated Channel Features. We proposed several approaches for the fusion process of the MACS hand position guess and the candidates of the hand detector. AfM is the first approach to produce good tracking results on our benchmark consisting of more than 25,000 frames under challenging conditions, quantified by the usual detection values and deviations from the ground truth paths. AfM reaches better detection values than MACS on its own and reduces the deviations from the ground truth path by more than 50 %. Furthermore the AfM ALL variant moves the decision if a hand is visible from the particle filter of MACS to the fusion process by rejecting candidates of the hand detector by their score, through which AfM ALL yields much more generality.

Future work will concentrate on a better and more automated scene segmentation that does not need any assumptions about the number of motion layers or the size of the hand as still needed in MACS to make AfM more general. One should even think about an online training phase of the hand detector. Since AfM comes close to good tracking, the benchmark needs to be improved with further sequences containing other moving objects and more people performing the gestures in different scenes.

References

[1] Oikonomidis, I., Lourakis, M.I.A., Argyros, A.A.: Evolutionary quasi-random search for hand articulations tracking. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3422–3429. CVPR '14, IEEE Computer Society, Washington, DC, USA (2014)

[2] Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and Robust Hand Tracking from Depth. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1106–1113 (Jun 2014)

[3] de La Gorce, M., Fleet, D., Paragios, N.: Model-Based 3d Hand Pose Estimation from Monocular Video. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(9), 1793–1805 (Sep 2011)

[4] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision 92, 1–31 (2011)

[5] Li, C., Kitani, K.M.: Pixel-level hand detection in ego-centric videos. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 3570–3577. IEEE (2013)

[6] Pisharady, P., Vadakkepat, P., Loh, A.: Attention based detection and recognition of hand postures against complex backgrounds. International Journal of Computer Vision 101, 403–419 (2013)

[7] Hammer, J. H., Beyerer, J.: Robust hand tracking in realtime using a single head-mounted rgb camera. In: Kurosu, M. (ed.) Human-Computer Interaction. Interaction Modalities and Techniques, Lecture

Notes in Computer Science, vol. 8007, pp. 252–261. Springer Berlin Heidelberg (2013)

[8] Hammer, J. H., Voit, M., Beyerer, J.: Motion Segmentation and Appearance Change Detection Based 2D Hand Tracking. In: To appear at the 19th International Conference on Information Fusion (July 2016).

[9] Wang, R.Y., Popović, J.: Real-time hand-tracking with a color glove. ACM Trans. Graph. 28(3), 63:1–63:8 (Jul 2009)

[10] Mistry, P., Maes, P.: Sixthsense: a wearable gestural interface. In: ACM SIGGRAPH ASIA 2009 Sketches. pp. 11:1–11:1. SIGGRAPH ASIA '09, ACM, New York, NY, USA (2009)

[11] Prisacariu, V., Reid, I.: Robust 3d hand tracking for human computer interaction. In: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. pp. 368 –375 (march 2011)

[12] Appenrodt, J., Al-Hamadi, A., Elmezain, M., Michaelis, B.: Data gathering for gesture recognition systems based on mono color-, stereo color- and thermal cameras. In: Proceedings of the 1st International Conference on Future Generation Information Technology. pp. 78–86. FGIT '09, Springer-Verlag, Berlin, Heidelberg (2009)

[13] Bradski, G.R.: Real time face and object tracking as a component of a perceptual user interface. In: Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (WACV'98). pp. 214–. WACV '98, IEEE Computer Society, Washington, DC, USA (1998)

[14] Kölsch, M., Turk, M.: Fast 2d hand tracking with flocks of features and multi-cue integration. In: Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on. p. 158 (june 2004)

[15] Phung, S., Bouzerdoum, A., S., Chai, D., S.: Skin segmentation using color pixel classification: analysis and comparison. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(1), 148 –154 (jan 2005)

[16] Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. Pattern Recognition 40(3), 1106 – 1122 (2007)

[17] Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: Proceedings of the 29th DAGM conference on Pattern recognition. pp. 214–223. Springer-Verlag, Berlin, Heidelberg (2007)

[18] Needham, C.J., Boyle, R.D.: Performance evaluation metrics and statistics for positional tracker evaluation. In: Proceedings of the 3rd international conference on Computer vision systems. pp. 278–289. ICVS'03, Springer-Verlag, Berlin, Heidelberg (2003)

[19] DHL Customer Solutions & Innovation. Augmented Reality in Logistics. Changing the way we see logistics - a DHL perspective. http://www.dhl.com/content/dam/downloads/g0/about_us/logistics_insights/csi_augmented_reality_report_290414.pdf, 2014, Accessed: 2016-03-14.

[20] Nils Petersen and Didier Stricker. Morphing billboards: an image-based appearance model for hand tracking. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–13, January 2014.

[21] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. Accurate, Robust, and Flexible Real-time Hand Tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3633–3642, New York, NY, USA, 2015. ACM.

[22] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2015.

[23] Deqing Sun, E.B. Sudderth, and M.J. Black. Layered segmentation and optical flow estimation over time. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1768–1775, 2012.

[24] Brian Taylor, Vasiliy Karasev, and Stefano Soatto. Causal Video Object Segmentation From Persistence of Occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4268–4276, 2015.

[25] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, August 2014.

[26] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, pp. 886-893 vol. 1.