Robust Statistics for Feature-based Active Appearance Models

Marcin Kopaczka, Philipp Gräbel and Dorit Merhof

Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany

Keywords: Active Appearance Models, Occlusion, Robust Cost Functions, Feature Descriptors.

Abstract: Active Appearance Models (AAM) are a well-established method for facial landmark detection and face tracking. Due to their widespread use, several additions to the original AAM algorithms have been proposed in recent years. Two previously proposed improvements that address different shortcomings are using robust statistics for occlusion handling and adding feature descriptors for improved landmark fitting performance. In this paper, we show that a combination of both methods is possible and provide a feasible and effective way to improve robustness and precision of the AAM fitting process. We describe how robust cost functions can be incorporated into the feature-based fitting procedure and evaluate our approach. We apply our method to the challenging 300-videos-in-the-wild dataset and show that our approach allows robust face tracking even under severe occlusions.

1 INTRODUCTION

Active Appearance Models (AAMs) are a state-of-the-art method for facial landmark detection and tracking. Introduced by Cootes in 1998 (Cootes et al., 1998), they have ever since been an active area of research and several relevant improvements to the original algorithm have been proposed in recent years. Apart from publications focusing on understanding and improving the AAM fitting process that will be described in more detail in Section 2.1, research has also focused on improving robustness towards challenging in-the-wild scenarios, unseen faces and occlusions. In this paper, we will focus on two of the proposed approaches:

- Applying methods from robust statistics to AAM fitting: as shown in (Gross et al., 2006; Theobald et al., 2006), outlier-resistant robust statistics can be used to improve AAM fitting under the presence of occlusions. Robust statistics allow analyzing modeled faces to detect occlusions and allow a subsequent exclusion of occluded face areas from the computation of the optimizer's error value. This way, only face areas that are more likely to be occlusion-free will contribute to the computation of the final landmarks.
- Feature-based AAMs as introduced in (Antonakos et al., 2015; Antonakos et al., 2014): instead of using the input image itself for fitting, feature-based AAMs increase the number of image channels by applying a defined feature descriptor such as HOG (Dalal and Triggs, 2005) or dense SIFT (Lowe, 1999) densely to each image pixel and performing the fitting on the resulting multichannel feature image. This approach allows incorporation of local neighborhood information into the holistic AAM algorithm, thereby increasing precision and robustness of the fitting process.

Each of the above methods has been shown to improve AAM fitting performance by addressing different aspects of the original algorithm. In our work, we will therefore inspect methods for combining both approaches and show on a challenging set of artificially occluded real-world videos from the 300-videos-in-the-wild dataset that our combined approach allows better handling of occluded video sequences than unmodified AAMs or applying each of the methods alone.

The structure of this paper is as follows: The next section will describe regular and intensity-based AAMs and robust cost functions, while Section 3 introduces our approach to combining robust cost functions with intensity-based multichannel AAMs. Section 4 describes a number of tests conducted to evaluate these algorithms as well as the results.

2 PREVIOUS WORK

Here, we give a brief overview over constructing and fitting an active appearance model, furthermore feature-based AAMs and robust statistics for occlusion handling will be described.

Kopaczka, M., Gräbel, P. and Merhof, D.
Robust Statistics for Feature-based Active Appearance Models.
DOI: 10.5220/0006910004210426
In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018) - Volume 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS*, pages 421-426
ISBN: 978-989-758-319-3

2.1 Active Appearance Models

Active Appearance Models have been introduced by Cootes et al. (Cootes et al., 2001) and significantly improved in their applicability by Matthews and Baker (Matthews and Baker, 2004) by introducing the project-out inverse compositional algorithm. An AAM is built using a database of annotated training images which are aligned via procrustes analysis, undergo a principal component analysis and are subsequently split into a shape and an appearance model. The shape model defines landmark positions via a mean shape $\boldsymbol{s}_0$ and a fixed number of eigenvectors $\boldsymbol{S}$, which contain the most significant deviations from the mean shape. New shapes can be modelled using a parameter vector $\boldsymbol{p}$: $\boldsymbol{s}_p = \boldsymbol{s}_0 + \boldsymbol{S}\boldsymbol{p}$. The same principle applies to the appearance model which represents the images's texture component. Using a mean appearance $\boldsymbol{A}_0$ and eigenfaces $\boldsymbol{A}$ warped into the mean shape, a model instance is defined by a set of parameters $\boldsymbol{\lambda}$: $\boldsymbol{A}_\lambda = \boldsymbol{A}_0 + \boldsymbol{A}\boldsymbol{\lambda}$.

The process of finding appropriate model parameters that optimally represent a given face is called fitting and solved using multidimensional optimization algorithms. The most current and complete overview of fitting algorithms can be found in the 2016 work by Alabort-i-Medina et al. (Alabort-i-Medina and Zafeiriou, 2016). A precise yet computationally fast method described in their work is the Wiberg Inverse Compositional Algorithm, which iteratively computes shape parameter updates similar to a Gauss-Newton optimisation and appearance parameter with a closed solution. This aims to find a set of parameters that minimizes the cost function $c = ||\boldsymbol{I}(W(\boldsymbol{p})) - \boldsymbol{A}_\lambda||_2^2$, in which $\boldsymbol{I}(W(\boldsymbol{p}))$ describes the current image warped into the mean shape space using the parameter vector $\boldsymbol{p}$. The difference image is often called *residual*.

Feature-based AAMs have been proposed by Antonakos et al. in 2014 (Antonakos et al., 2014). They allow AAM fitting on a multichannel feature image computed from the original image using a feature descriptor. While increasing the computational load, including features into the AAM fitting process allows to enhance specific image properties such as local gradient or texture information depending on the used descriptor. Especially gradient-oriented features such as HOG or DSIFT have been shown to improve fitting performance. In combination with a diversified training database such as LFPW (Belhumeur et al., 2013), feature-based AAMs have shown a remarkable performance when fitting faces under challenging conditions such as strong illumination variation and partial occlusions.

2.2 Robust Statistics for AAM Fitting

As active appearance model fitting compares the intensities of a model – which in our assumption is trained on data containing only limited types of occlusion or no occlusion at all – and the original image, occlusions can have a large impact on fitting accuracy and may lead to a diverged result. This is even more problematic in video tracking, as a fitting result is commonly used as the initial shape estimate for the next frame, therefore potentially increasing misalignment with each subsequent frame. To allow detection of occluded pixels, Gross et al. (Gross et al., 2006) introduced the idea of applying robust cost functions to the images during model fitting. Robust cost functions weight each pixel of the warped image based on the value of the residual and this pixel's standard deviation in the training data set. The basic assumption is that high residual values are likely to be caused by occlusions in the original image. Depending on the used robust cost function, occluded pixels contribute only partially or not at all to the final fitting result. Different cost functions can be used to compute pixel weights; in (Theobald et al., 2006), several functions in terms of detection accuracy and fitting robustness have been inspected.

Our work presents a novel approach that combines feature-based models with robust cost functions. Since robust cost functions require an estimation of the standard deviation of residual values, we also introduce and evaluate methods to compute these for feature-based models.

3 METHODOLOGY

3.1 Standard Deviations

In this work, standard deviations are used to normalize residuals prior to the fitting procedure. Similar to (Gross et al., 2006), we use the median absolute deviation (MAD) (Huber and Ronchetti, 1981) as an estimate. The standard deviation is computed for each model pixel. In addition to using this method for computing pixel-wise MADs, this paper introduces the novel idea of using a single, image-wide MAD. The motivation behind this approach is our observation from initial experiments that have shown that MAD values computed on the model's DSIFT channels are very similar across all channels. This leads to the following two approaches to compute standard deviations and normalize residuals which will both be inspected in this paper: **Pixel-wise** computation as introduced in (Gross et al., 2006) is based on the residuals of the last iteration, and computes the MAD for every single pixel in every single channel (intensity and/or one of the 36 DSIFT channels). This allows the standard deviation estimates to be dependent on the position in the image, reflecting that some areas such as the cheeks may be more homogeneous than other high-variance areas such as mouth or eyes. Additionally, we introduce an **image-wide** method that computes a combined MAD for all channels based on the results of the last iteration. The 36 DSIFT channels are analyzed in combination, resulting in one value for the intensity channel and equal values for every DSIFT channel. This approach allows a computationally inexpensive MAD computation. Areas with higher standard deviation estimates in the pixel-wise case – often detailed facial features – are more likely described as occluded in this case.

3.2 Applying Robust Cost Functions to Feature-based AAMs

Instead of performing Gauss-Newton optimization $\Delta \boldsymbol{p} = \boldsymbol{H}^{-1}\boldsymbol{J}^T\boldsymbol{r}$, where $\boldsymbol{H} = \boldsymbol{J}^T\boldsymbol{J}$ is an approximation of the Hessian and $\boldsymbol{J}$ the Jacobian of the residuals with respect to the parameter vector $\boldsymbol{p}$, a weight map $\boldsymbol{W}$ is applied according to the iteratively re-weighted least squares algorithm: $\Delta \boldsymbol{p} = \boldsymbol{H}_W^{-1}\boldsymbol{J}_W^T\boldsymbol{r}$ with $\boldsymbol{H}_W = \boldsymbol{J}^T\boldsymbol{W}\boldsymbol{J}$ and $\boldsymbol{J}_W = \boldsymbol{W}\boldsymbol{J}$. $\boldsymbol{W}$ is a diagonal matrix containing the weights w_i, which are computed according to the robust cost functions in (Theobald et al., 2006). This work uses the Huber, Talwar, Tukey bisquare and Cauchy robust cost functions as well as standardized distance (stadis), Gaussian probability density function (pdf) and decaying exponentials (decexp) to compute weights:

$$w_{\text{huber}}(r_i) = \begin{cases} 1 & \text{if } |r_i| \leq k \\ \frac{k}{|r_i|} & \text{if } |r_i| > k \end{cases} \tag{1}$$

$$k_{\text{huber}} = 1.345\,\sigma \tag{2}$$

$$w_{\text{talwar}}(r_i) = \begin{cases} 1 & \text{if } |r_i| \leq k \\ 0 & \text{if } |r_i| > k \end{cases} \tag{3}$$

$$k_{\text{talwar}} = 2.795\,\sigma \tag{4}$$

$$w_{\text{bisquare}}(r_i) = \begin{cases} \left(1 - \frac{r_i^2}{k^2}\right)^2 & \text{if } |r_i| \leq k \\ 0 & \text{if } |r_i| > k \end{cases} \tag{5}$$

$$k_{\text{bisquare}} = 4.685\,\sigma \tag{6}$$

$$w_{\text{cauchy}}(r_i) = \frac{1}{1 + \left(\frac{|r_i|}{k}\right)^2} \tag{7}$$

$$k_{\text{cauchy}} = 2.385\,\sigma \tag{8}$$

$$w_{\text{stadis}}(r_i) = \begin{cases} 1 & \text{if } \frac{|r_i|}{\sigma} \leq 2\sigma \\ 0 & \text{if } \frac{|r_i|}{\sigma} > 2\sigma \end{cases} \tag{9}$$

$$w_{\text{pdf}}(r_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|r_i|}{2\sigma^2}} \tag{10}$$

$$w_{\text{decexp}}(r_i) = e^{-\frac{|r_i|}{2\sigma^2}} \tag{11}$$

Since both the original image data (often called the intensity channel) and the DSIFT channels contain complementary information we extend the DSIFT model with an additional channel containing the unprocessed image itself. The values of both intensity and DSIFT descriptor are weighted according to the chosen estimate of standard deviations. We propose and evaluate two approaches for weight map computation on feature-based multichannel models: The weight can be computed either independently for each channel or on the intensity channel only and then applied to the DSIFT channels. The motivation behind using the intensity channel for map computation is computation speed as well as the assumption that information on potential occlusions can be found preferably in this channel.

4 EXPERIMENTS AND RESULTS

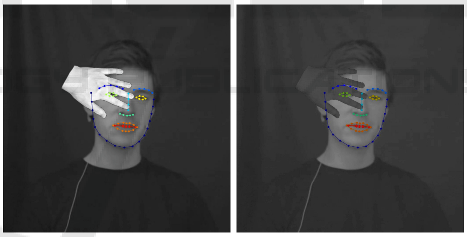

Figure 1: One artificially occluded frame (left: light hand, right: dark hand) with 18.5 % occlusion and 14 of 68 landmarks covered. As the occlusion is added algorithmically, the ground truth landmarks are known.

Our method was evaluated on a subset of the *300 Videos in the Wild* (Shen et al., 2015; G. S. Chrysos and Snape, 2015; Tzimiropoulos, 2015) test set. A ten second sequence from each video was selected and overlaid with a hand image which was either light or dark to compensate differences in skin tone. In every subsequence, the hand moved four times along the chin landmarks – each time occluding a larger part of the face. This results in a test set with 18352 frames in $2 \cdot 31$ sequences. The largest occluded area in the test sequences covers 69.1 % of the mean shape and 53 out of 68 landmark points.

The AAM was created using an AAM based on the Helen and LFPW training data – with consistent and corrected annotations provided by the iBug group (Sagonas et al., 2016). The combined model has a resolution of 120 pixels, two layers, 50 appearance and 9 shape components.

The error norm is the normalized mean Euclidean distance between ground truth shape and fitting result. The normalization factor is the average of height and width of the ground truth shape.

4.1 Computing the Standard Deviation

In an initial step, we have conducted preliminary experiments to determine the influence of the training data for the computation of standard deviation. To this end, we constructed active appearance models based on training data of the the Helen dataset, LFPW, and both. For each model, the corresponding training as well as test data was used to compute image-wide median absolute deviation values. Results suggest that that MADs – and therefore, standard deviations – do not depend significantly on the input images used to train the model. Therefore, the following evaluations are based on the standard deviations estimated based on the combined Helen and LFPW training sets as this was the combination containing the maximal possible amount of image data.

4.2 Experiments

Two main experiments have been conducted to evaluate our methods: For the first set, each of the cost functions described in 3.2 has been evaluated on each of the test sequences to determine the best performing cost function. Image-wide standard deviation estimates have been used with an intensity only weight map calculation. As baseline, we also track the sequences with an AAM using the same features (intensity plus DSIFT with normalization by standard deviation) whithout applying the robust cost function. For the second set of experiments, using the probability density function as weights for the iteratively reweighted least squares algorithm, every combination of standard deviation ('image-wide' and 'pixel-wise') and weight map computation ('all channels' and 'intensity only') was evaluated. For comparison, a regular active appearance model was fitted as well. For the regular model, the variants 'intensity only' and 'all channels' are equivalent since the model consists of a single intensity channel. Every sequence was tracked independently and initialized with a perturbed shape from the ground truth's bounding box. The fitting result of each frame was then subsequently used as starting position for the succeding frame.

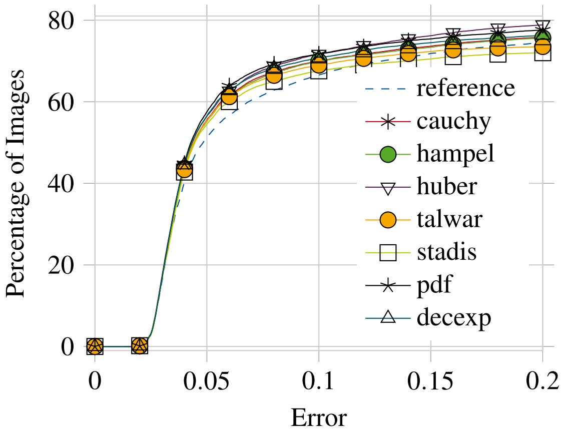

Figure 2: Cumulative error distribution plot of cost function performance.

4.3 Results

All results are plotted as a cumulative error distribution of the error norm values of all frames, thereby allowing evaluation of fitting accuracy as well as robustness. Steeper curves indicate better fitting performance. The results of the cost function evaluation can be found in Figure 2. They show that using our method for combining robust cost functions with feature-based AAMs improves both accuracy and robustness of the fitting procedure. Out of all evaluated cost functions, only the standardized distance and the Talwar threshold yield a slightly lower robustness, while still being more precise for the most part. It is interesting to note that these two are the only binary weight functions (either 0 or 1, no values in between) that have been tested, while all other approaches yield continous results between 0 and 1. The best performing weight functions are the Huber function and the probability density function, which also performed best in (Theobald et al., 2006) for robust fitting of regular AAMs. The results for the evaluation of our novel standard deviation and weight map computation methods are displayed in Figure 3. It shows that using feature-based models with robust cost functions outperforms the robust intensity-based model fitting introduced in (Gross et al., 2006) in all cases. The significant differences are caused by evaluating the methods on image sequences and not on single frames: Errors propagate from one frame to the next and may lead to larger deviations over time. Our experiments show that it is generally beneficial to use only the intensity channel for computation of a weight map, even if a robust feature descriptor is used. This might be due to DSIFT using neighborhood information and not single pixels, thereby lowering the spatial resolution of the weight map. The impact of choosing

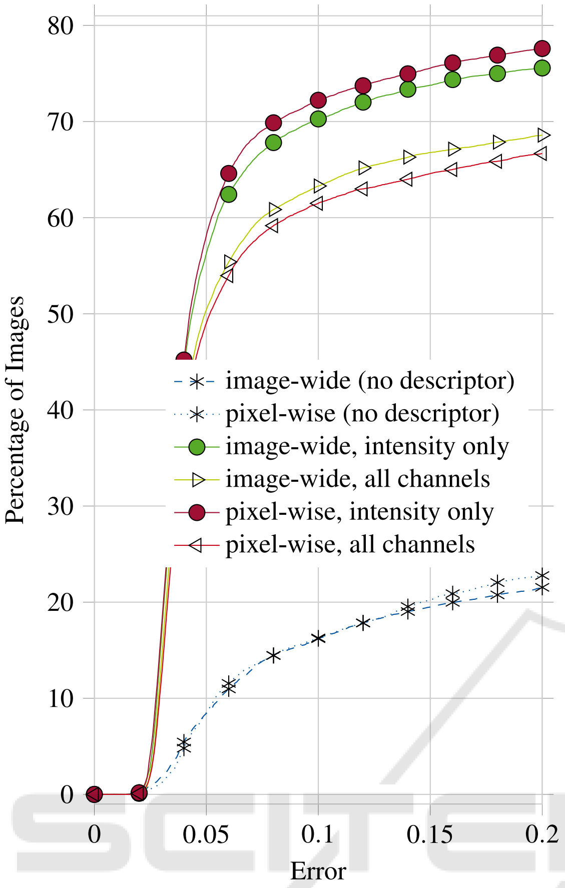

Figure 3: Cumulative error distribution plot comparing feature-based and regular AAMs, weight maps computed on all channels or on the intensity channel only and global vs. per-pixel intensity maps.

either image-wide or pixel-wise intensity map computation on the fitting performance is small. Therefore, image-wide intensity map computation is a viable approach especially when additionally considering its drastically reduced computation time.

5 CONCLUSION

We have shown that AAM fitting precision can be improved by combining robust cost functions with feature-based AAMs. We presented and evaluated two methods to estimate the standard deviations and two methods to compute weight maps that are both needed for robust fitting. An extensive evaluation has shown that our proposed approach allows highly improved tracking performance in video sequences compared to regular AAMs.

REFERENCES

Alabort-i-Medina, J. and Zafeiriou, S. (2016). A Unified Framework for Compositional Fitting of Active Appearance Models. *International Journal of Computer Vision*, pages 1–39.

Antonakos, E., Alabort-i-Medina, J., Tzimiropoulos, G., and Zafeiriou, S. (2014). HOG Active Appearance Models. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 224–228.

Antonakos, E., Alabort-i-Medina, J., Tzimiropoulos, G., and Zafeiriou, S. (2015). Feature-Based Lucas-Kanade and Active Appearance Models. *Image Processing, IEEE Transactions on*, 24:2617–2632.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). Active appearance models. In *European conference on computer vision*, pages 484–498. Springer.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active Appearance Models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 681–685.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

G. S. Chrysos, E. Antonakos, S. Z. and Snape, P. (2015). Offline deformable face tracking in arbitrary videos. *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Gross, R., Matthews, I., and Baker, S. (2006). Active Appearance Models with Occlusion. *Image and Vision Computing*, 24:593–604.

Huber, P. and Ronchetti, E. (1981). *Robust Statistics*. Wiley and Sons.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee.

Matthews, I. and Baker, S. (2004). Active Appearance Models Revisited. *International Journal of Computer Vision*, 60:135–164.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18.

Shen, J., Zafeiriou, S., Chrysos, G. S., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Theobald, B.-J., Matthews, I., and Baker, S. (2006). Evaluating Error Functions for Robust Active Appearance Models. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 149–154. IEEE.

Tzimiropoulos, G. (2015). Project-out cascaded regression with an application to face alignment. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659—3667.