ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1812

A Practical Approach to Fuse Shape and Appearance Information in a Gaussian Facial Action Estimation Framework

Teena Hassan[1] and **Dominik Seuss**[1] and **Johannes Wollenberg**[1] and **Jens Garbas**[1] and **Ute Schmid**[2]

Abstract. In many domains of computer vision, such as medical imaging and facial image analysis, it is necessary to combine shape (geometric) and appearance (texture) information. In this paper, we describe a method for combining geometric and texture-based evidence for facial actions within a Kalman filter framework. The geometric evidence is provided by a face alignment method. The texture-based evidence is provided by a set of Support Vector Machines (SVM) for various Action Units (AU). The proposed method is a practical solution to the problem of fusing categorical probabilities within a Kalman filter based state estimation framework. A first performance evaluation on upper face AUs demonstrates the practical applicability of the proposed fusion method. The method is applicable to arbitrary imaging domains, apart from facial action estimation.

1 INTRODUCTION

Two important types of information commonly used in computer vision are shape and appearance. For example, in medical imaging, shape and appearance information have been used for automatic classification of cells [18]. In this paper, we focus on combining both sources of information for facial expression analysis. Facial expression is accompanied by deformation of the shape of facial features and by changes in facial texture. Changes in shape of facial features–such as eyes, nose, eyebrows, or lips–constitute geometric information. Changes in facial appearance–such as wrinkles, furrows, edges, or bulges–constitute texture information.

The emergence of robust face detection algorithms [31] in early 2000s accelerated research on automatic analysis of faces recorded in images and videos. Automatic analysis of facial expressions is one of the research fields that received increased attention since then [27, 35, 10]. Research in this field is pursued mainly in two directions [35]: One of the directions focuses on an objective analysis of basic facial movements based on the Facial Action Coding System (FACS) [9]. The other focuses on detecting a set of prototypical facial expressions of emotions.

Since more than ten years, the Intelligent Systems group at Fraunhofer IIS has been developing the Sophisticated High-speed Object Recognition Engine SHORE™[3], which is a general framework for various object detection tasks in images and videos, with a focus on face detection and analysis [25]. SHORE™ detects and tracks faces in real-time [20], estimates age and gender, and identifies four basic expressions of emotions, namely happiness, sadness, anger and surprise.

In collaboration with the GfK Verein (a think tank for market research and majority shareholder of Europe's largest market research company GfK SE), the software EMOScan was developed by customising and extending SHORE™ for valence detection [12]. EMOScan is currently deployed in nine European, three Asian and two American countries. It is used in online and offline studies for analysing viewers' facial responses to advertisements. Currently, SHORE™ is also being applied to several other application areas of automatic facial expression analysis, such as affective computing, pain detection and autism research.

While the focus of SHORE™ was primarily on emotion detection, our current research focuses on FACS-based analysis, where image sequences are processed for detecting facial motion and coding it in terms of FACS Action Units (AU). This approach enables detection of more subtle facial expressions. The detected AUs and their intensities could be used in a subsequent processing step for a detailed analysis of human affective behaviour. This is planned as an extension of SHORE™ with a similar deployment strategy.

In the next section, previous research on the fusion of geometric and texture information in the domain of facial expression analysis is reviewed. Afterwards, an overview of the main components of our AU estimation framework is provided. This is followed by a detailed description of the proposed fusion approach within a Gaussian state estimation framework. We present a performance evaluation of the fusion approach based on a dataset collected in a private study conducted by GfK Verein. Finally, we conclude with an outlook to future work.

2 RELATED WORK

Recently, fusion of shape (geometric) and appearance (texture) features has shown to give promising results in the field of AU detection and emotion recognition. Geometric features are usually computed using the location of facial landmarks defined according to a deformable face model. Typical approaches for landmark localization include Active Appearance Models (AAM) [5] and Constrained Local Model Fitting (CLM) [26]. Texture features encode visual texture information using, for example, histograms of oriented gradients (HOG) [6], histograms of local binary patterns (LBP) [13] or histograms of local Gabor binary patterns (LGBP) [2].

When it comes to fusion of geometric and texture features, machine learning approaches for facial expression analysis have em-

[1] Fraunhofer Institute for Integrated Circuits IIS, Germany, email: teena.hassan@iis.fraunhofer.de
[2] Faculty of Information Systems and Applied Computer Science, Otto-Friedrich-Universität Bamberg, Germany
[3] http://www.iis.fraunhofer.de/shore

ployed different strategies, namely, feature-level fusion, decision-level fusion, and fusion based on multiple kernel learning. Feature-level fusion is usually done by concatenation of all features into one large feature vector, which is then passed to a machine learning algorithm [3, 34, 33, 23, 36]. Decision-level fusion usually trains separate classifiers with geometric or texture features, and later combines their scores. For example, [17] uses a variant of artificial neural networks, and [16] uses an SVM to fuse scores from classifiers. Multiple kernel learning based fusion approaches use separate kernels for geometric and texture features. In [22], a multi-kernel SVM is used for feature fusion. A Gaussian kernel is used for geometric and gradient-based texture features, and an intersection kernel is used for higher-dimensional Gabor-based texture features. A multi-kernel SVM is also used in [28] for feature fusion, where kernels of the same type are applied to geometric and texture features.

Fusion of geometric and texture information can also be viewed as a sensor fusion problem. Traditionally, state estimation methods are used to perform sensor fusion, for example, in applications like robot navigation and satellite tracking. State estimation methods fuse a stochastic model of the temporal dynamics of the state of a system, and the stochastically distributed measurements provided by multiple sources. Dynamic state estimation methods such as Kalman filtering, particle filtering, Hidden Markov Models (HMM), and Dynamic Bayesian Networks (DBN) have been used for facial expression analysis [29, 8, 14, 37], in order to estimate intensities of AUs [8, 14, 7] and facial expressions of emotions [8, 37], as well as to infer the temporal phases of AUs [29]–such as onset, apex and offset.

AU intensity estimation is a continuous state-space estimation problem. Therefore, Kalman and particle filtering based approaches–such as [8, 14]–have been applied. These methods commonly use deformable face models that encode the semantics of the different facial expressions to be estimated, and use physics-based models to describe the temporal dynamics of facial motion. The parameters of pre-defined temporal models are learned from annotated data. Geometric or texture features constitute the measurements. These are fused by defining a measurement model that provides the likelihood of observing the geometric or texture features, given the current state estimate. The advantages of such approaches are their lower data requirements, and their simple and intuitive way of combining semantic, spatial and temporal aspects of facial motion.

However, the performance of these state estimation methods depends on the accuracy of the models and the correctness of the stochastic assumptions. For example, precise modelling of all possible variations in the shape, appearance and dynamics of facial expressions is not possible. To overcome these limitations, outputs from data-driven machine learning approaches could be used to improve the discriminative power of state estimation based approaches [24, 29, 30, 19]. The integeration of probability outputs from SVMs in an HMM-based discrete state estimation framework has been explored in application domains such as speech recognition [19] and recognition of temporal phases of AUs [29, 30]. However, their integration in a continuous, Gaussian-distributed state estimation framework appears to be unexplored.

The use of a state estimation method allows the possibility of filtering noisy measurements and enables the use of a dynamic model that approximates the mechanics of facial muscles. It provides a temporal smoothing effect on the output estimates that is usually lacking in machine learning approaches. Temporally smooth estimates improve the chances of finding good thresholds for discretisation, if needed. State estimation frameworks also provide an inherent mechanism to integrate information from multiple modalities or sources. In the light of these arguments, we use a state estimation based approach for estimating continuous-valued AU intensities.

We use anatomically-inspired models to capture the relevant spatio-temporal aspects of facial motion, such as the facial deformations resulting from the motion and the biomechanics of facial muscles that produce the motion. Since the state space is high dimensional, a Kalman filter based framework is used. To improve the discriminative capacity of the system, we use SVMs that are trained on texture features to discriminate between two or more AU classes. The Kalman filter assumes the state and measurements to be Gaussian-distributed. This gives rise to the problem of integrating categorical probabilities from SVMs into a Gaussian-distributed continuous state estimation framework. In this paper, we propose a practical approach to solve this problem. The applicability of the proposed approach is demonstrated within our AU estimation framework.

3 SYSTEM OVERVIEW

Estimating action unit intensities under real conditions, without the need for explicitly adapting the system to the monitored person, is very challenging. Our system tackles variations in lighting, head pose and interpersonal facial shape with the help of robust face detection, face normalisation and an online face calibration. The flowchart of our system is shown in Figure 1.

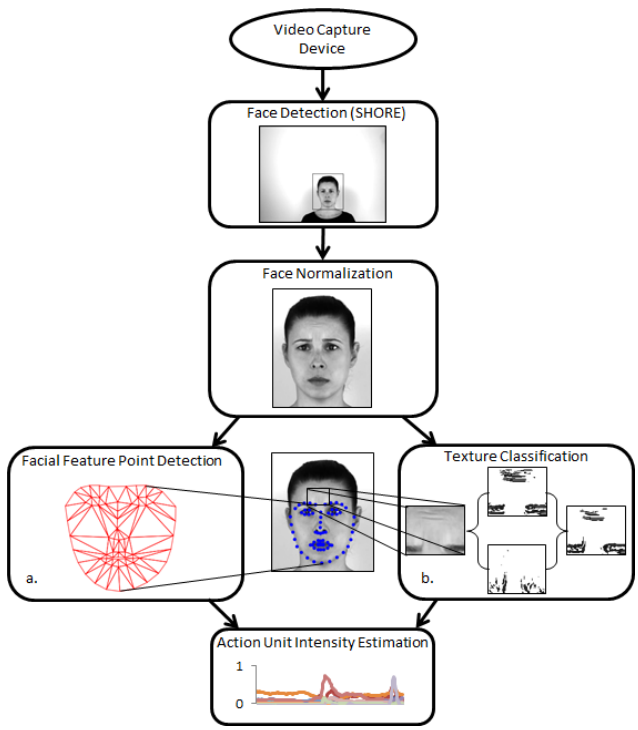

Figure 1. Flow chart of the system. The extracted face mesh is shown in (a). Exemplary texture features extracted from the forehead region using vertical and horizontal edge filters are shown in (b).

- **Face Detection:** The face detection unit processes the image frames provided by a video capture device. It uses SHORE™ [20] to locate the person's face in the image and obtain the region of the face. In addition, SHORE™ provides the location of eyes, nose and mouth corners, in case a face could be found. If more than one

face is present, the face detector selects the most prominent face on the basis of the face size in the image.

- **Face Normalization:** The face normalization unit rotates and scales the person's face using the five facial feature points provided by SHORE™ as reference points. Thus, the normalized image always has the same resolution and pose. In this way, some of the variations in the appearance of the face that are caused by head rotations and movements of the person in front of the video capture device, are mitigated.
- **Facial Feature Point Detection:** This unit determines the location of additional facial feature points within the detected face region (Figure 1.a). These points (face mesh) cover the whole face and track the location of prominent spots on the human face, such as lip boundaries, eye brows, chin, etc. Changes in the location of these points from one frame to another provide geometric information about facial motion and the activated AUs.
- **Texture Classification:** However, it is not possible to detect all AUs just by observing motion in these facial feature points, since some of them are more prominently expressed by wrinkles. Such AUs are easier to recognise from the transient changes in the appearance of the face; for example, the combination of AU01 (Inner Brow Raiser) and AU04 (Brow Lowerer) is easily recognised from the appearance of wrinkles on the forehead region (Figure 1.b). So, in addition to the facial feature points, the facial texture is analysed. SVM classifiers trained on texture features are used for detecting specific AUs and AU combinations.
- **Action Unit Intensity Estimation:** The Kalman filter based action unit intensity estimator fuses the outputs from the facial feature point detector and the texture classification unit to get a final estimate of the intensity of each AU in a pre-defined set of 22 AUs. A biomechanical model of facial muscles is used to model the dynamics of the AUs. The intensities of AUs are modelled as the parameters of an anatomically based deformable face model. In addition to estimating AU intensities, this unit also simultaneously performs an online dynamic calibration of the person's neutral face. This online calibration is neccesary because it is often not possible to acquire a neutral face on demand.

4 FUSION APPROACH

In our system, we fuse observations (measurements) from two types of sources, namely geometric and texture, within the Kalman filter framework to estimate the intensities of various AUs. The geometric measurements are the positions of facial feature points and the texture measurements are the class-wise success probabilities from SVMs trained on texture features extracted from the image.

The Kalman filter [15] is a special form of dynamic Bayesian network that is applied to continuous state spaces with Gaussian transition and observation probability density functions. Kalman filtering involves two steps: *predict* and *update*. In the *predict* step, a dynamic process model is applied to the previous state to predict the current state. The predicted estimate is called the *apriori* state estimate. In the *update* step, one or more measurements are used to correct the *apriori* state to obtain the filtered or *aposteriori* state estimate. The noise in the measurements are assumed to follow the (multivariate) zero-mean Gaussian distribution.

In the *update* step, the Kalman filter allows to fuse measurements from multiple sources, provided each source has a Gaussian noise model. The fusion is performed on the basis of the uncertainties in the measurements. The more reliable measurements contribute more to the state update than the less reliable measurements. To incorporate a measurement into the Kalman filter, two components are required:

1. A measurement model that maps the state variables to the measured variables.
2. A covariance matrix that describes the Gaussian noise in the measured variables.

The following subsections describe how to model these two components for the geometric and texture measurements used in our system.

4.1 Geometric measurements

The geometric measurements include the positions of 68 facial feature points detected by the facial feature point detection unit. These are detected using a face alignment method. The measurement model for geometric measurements is given by a 3D 68-point deformable shape model that is similar to the CANDIDE face model [1]. The covariance matrix for the noise associated with the geometric measurements is determined empirically by applying the face alignment method to an annotated dataset, such as the Extended Cohn-Kanade dataset [21]. The differences between the detected positions of facial feature points and their annotated positions are normalised relative to the annotated eye distance. The variances and covariances of these normalised errors constitute the elements of the covariance matrix. The noise in each point measurement could be assumed to be independent of that in others. In this case, the matrix would be block-diagonal. Alternatively, the noise in each point measurement could be assumed to be correlated to that in every other. In this case, the matrix would be a full matrix. At runtime, the covariance matrix is scaled by the square of the measured eye distance.

4.2 Texture measurements

The texture measurements include the probability outputs for different AU classes provided by SVMs trained on texture (appearance-based) features such as HOG or LBP. These measurements are provided by the texture classification unit of the system. The open source software library LIBSVM [4] is used to realise the texture classifiers. The probabilities of AU classes are interpreted as the intensities of corresponding AUs. This is under the assumption that the higher the intensity of an AU, the stronger the textural changes and the greater the class probability, and viceversa. Therefore, the measurement models for texture measurements are the identity functions involving the corresponding AU intensity parameters. This directly maps the probability of an AU class to its intensity of expression.

The probability of an AU class is determined in different ways, depending on the output configuration of SVM (two-class or multi-class). LIBSVM provides the probability of each SVM output class using the method based on pairwise coupling [32]. These are converted into AU class probabilities according to the output class definitions. Three cases are discussed below.

- **Case A:** A two-class SVM that detects the presence or absence of an AU A. This is the simplest case, where the probability output for class A provided by the SVM is used as-is.
- **Case B:** A multiclass SVM that detects all possible boolean combinations of occurrence of two or more AUs. In Table 1, an example involving two AUs A and B is given. In such cases, the probability of each AU can be obtained through marginalisation. From Table 1, the probability of A is computed as $p + q$, and that of B is computed as $p + r$.

- **Case C:** A multiclass SVM that detects several individual AUs. For example, a four-class SVM for A, B, C and rest. The probability of A is the output of the SVM for class A. The probability of absence of A is the sum of the probabilities for the other three classes. This is a generalisation of Case A.

Table 1. Table listing the four boolean combinations of occurrence of two AUs: A and B, and notations for the probability outputs from a corresponding four-class SVM.

Boolean Combinations	Probability Notations
A and B	p
A and not B	q
not A and B	r
not A and not B	s

p, q, r and s add to 1 (exhaustive)

The probabilities of AU classes so computed define a Bernoulli distribution for individual AUs. For a Bernoulli distributed AU variable A, the outcome '1' indicates the presence of A and the outcome '0' indicates the absence of A. If the probability of presence of A is p, then the first moment or expected value μ is computed as $\mu = 0(1-p) + 1(p) = p$. The expected value is therefore identical to the probability of presence of A. The second moment or variance σ^2 is given by $p(1-p)$. Figure 2 illustrates how the variance varies according to the probability p (a.k.a probability of success). As the probability approaches the extremities 0 or 1, the variance decreases, which indicates increasing confidence in the texture measurements. As the probablity approaches 0.5, the variance increases, which indicates decreasing confidence in the texture measurements. The skewness of the Bernoulli distribution of A is computed as $(1-2p)/\sqrt{p(1-p)}$. As a rule of thumb, normality could be assumed when the skewness varies between -2 and 2. This corresponds approximately to the probability range $[0.146, 0.854]$ for p. For simplicity and practical convenience, we assume normality throughout the probability range $[0, 1]$. Therefore, we use the second moment σ^2 of the Bernoulli distribution as the variance of the Gaussian noise associated with the texture measurement for A.

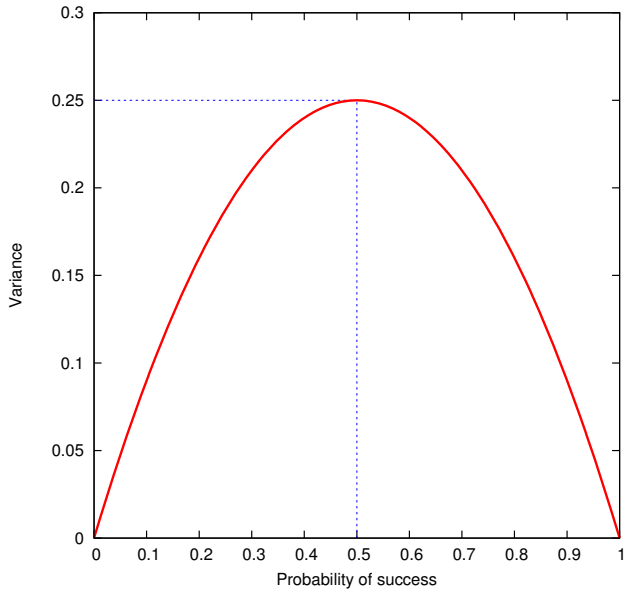

Figure 2. Variance of a Bernoulli distributed random variable

5 EVALUATION AND RESULTS

5.1 Dataset

We use an undisclosed dataset provided by GfK Verein for evaluation of the system and the proposed fusion approach. This dataset contains 301 videos of 93 different subjects (female and male). Several short advertisement clips were shown to the subjects and their facial reaction was recorded using a webcam, which was placed on top of the display screen. Each recording lasts approximately eight seconds and was recorded at 25 frames per sec. The subjects were recorded in a light-controlled setting and were instructed to act naturally. The responses often comprised of very subtle changes in facial expressions when exposed to the specific stimuli in the advertisement. Each frame is annotated by FACS experts with a list of active AUs.

5.2 Performance measure

We measure the performance of our system by creating the Receiver Operating Characteristic (ROC) curve [11] for each AU. ROCs illustrate the performance of a binary classifier system as its discrimination threshold is varied. The curve is a plot of the true positive rate against the false positive rate for various thresholds. Since ROCs are graphical plots, we compute the Area Under Curve (AUC) for each curve to get a single numerical metric. This makes comparison of performance of different configurations easier. A value for AUC that is closer to unity would indicate better performance.

5.3 Evaluation

We evaluate the proposed fusion approach on three upper face AUs, namely AU01 (Inner Brow Raiser), AU04 (Brow Lowerer) and AU06 (Cheek Raiser). As an illustration of **Case A** mentioned in Section 4.2, we use a pre-trained SVM for detecting the presence of AU06. As an illustration of **Case B** mentioned in Section 4.2, we use a pre-trained multiclass SVM for detecting the four possible Boolean combinations of AU01 and AU04 as mentioned in Table 1. **Case C** is a generalisation of **Case A**, and therefore, it is not separately evaluated. Tuning of system parameters as well as training and cross-validation of SVMs were performed using images selected from the CK+ dataset and an internal dataset of actors performing different AUs. The dataset provided by GfK Verein was not used for these purposes.

Table 2. AUC values for AU01, AU04 and AU06 for the fused, geometry-only and texture-only configurations.

Configuration	AU01	AU04	AU06
Fused	0.82	0.84	0.73
Geometry-only	0.77	0.79	0.64
Texture-only	0.78	0.79	0.66

In Table 2, the AUC values obtained after fusing geometric and texture information are provided. The corresponding ROC curves are shown in Figure 3. The improvement in performance obtained over (a) using only geometric information (geometry-only configuration),

and (b) using SVM probability outputs independently of Kalman filter framework (texture-only configuration), is illustrated in Figure 4. The evaluation shows that fusing geometric and texture information is better than using either of them alone. This indicates that, for AU01, AU04 and AU06, shape and appearance provide complementary information that helps in improving their recognition rates.

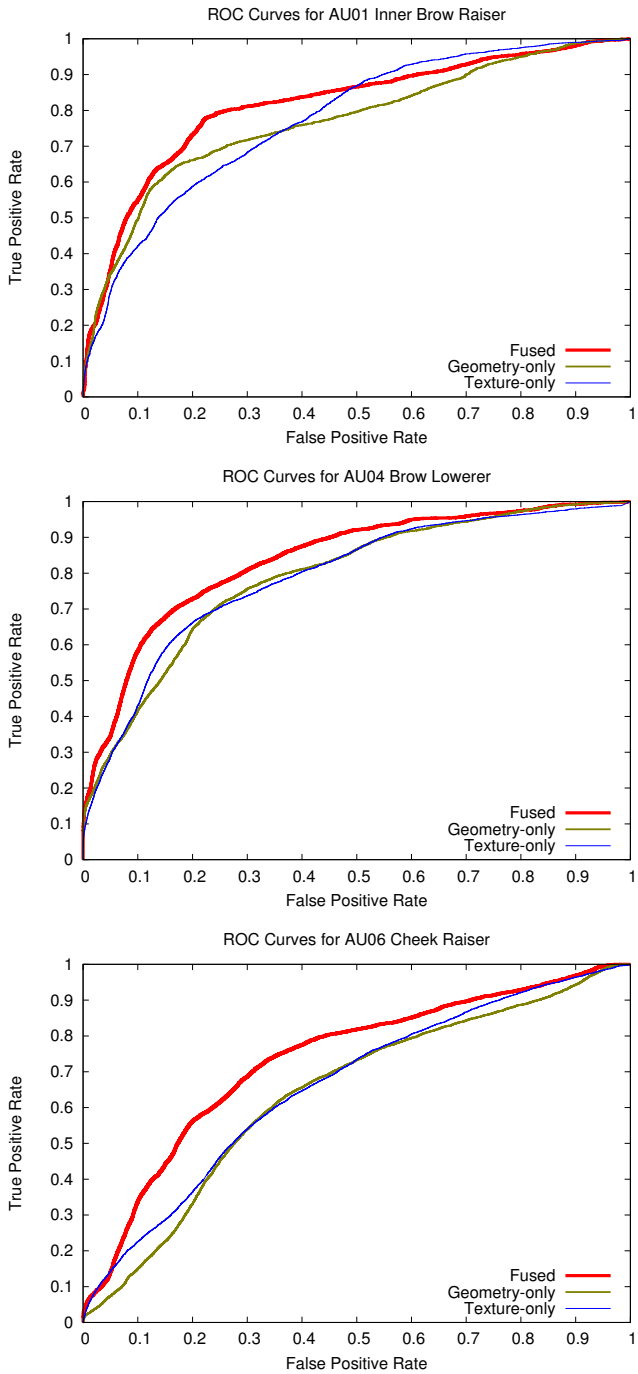

Figure 3. ROC curves for AU01, AU04 and AU06, for the fused, geometry-only and texture-only configurations.

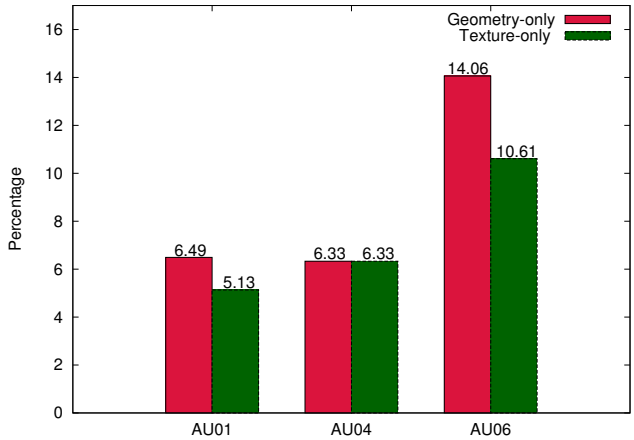

Figure 4. Percentage improvement in AUC of fused configuration over geometry-only and texture-only configurations for AU01, AU04 and AU06.

6 CONCLUSION

In this paper, we proposed a method to fuse categorical probabilities within a Kalman filter framework. This was illustrated in an application for facial action estimation by fusing geometric measurements of facial feature points with AU probabilities from SVM based texture classifiers. The results show that fusion of geometric and texture information using the proposed method clearly outperforms the geometry-only and texture-only configurations. This also illustrates the practical applicability of the proposed fusion method.

The proposed fusion method currently assumes normality across the entire range of probability output. However, the skewness of the distribution increases rapidly when the probability output approaches the extremities 0 and 1. Therefore, possible future work could include strategies for dealing with the skewness of the Bernoulli distribution of texture information.

ACKNOWLEDGEMENTS

We would like to thank the GfK Verein, in particular Anja Dieckmann and Matthias Unfried, for their support, valuable feedback and provision of the facial expression dataset. We would also like to thank Andreas Ernst and Sebastian Hettenkofer of Fraunhofer IIS for their feedback and active participation in discussions on the topic of this paper.

REFERENCES

[1] J. Ahlberg, 'CANDIDE-3 – an updated parameterized face', Technical Report LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden, (2001).

[2] T. R. Almaev and M. F. Valstar, 'Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition', in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 356–361, (Sept 2013).

[3] T. Baltrusaitis, M. Mahmoud, and P. Robinson, 'Cross-dataset learning and person-specific normalisation for automatic action unit detection', in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 06, pp. 1–6, (May 2015).

[4] Chih-Chung Chang and Chih-Jen Lin, 'LIBSVM: A library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27, (2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor, 'Active appearance models', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 681–685, (2001).

[6] N. Dalal and B. Triggs, 'Histograms of oriented gradients for human detection', in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 886–893 vol. 1, (June 2005).

[7] F. Dornaika and F. Davoine, 'On appearance based face and facial action tracking', *IEEE Transactions on Circuits and Systems for Video Technology*, **16**(9), 1107–1124, (Sept 2006).

[8] Fadi Dornaika and Franck Davoine, 'Simultaneous facial action tracking and expression recognition in the presence of head motion', *International Journal of Computer Vision*, **76**(3), 257–281, (2007).

[9] Paul Ekman and Wallace V. Friesen, *Facial action coding system: A technique for the measurement of facial movement*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[10] B. Fasel and Juergen Luettin, 'Automatic facial expression analysis: a survey', *Pattern Recognition*, **36**(1), 259 – 275, (2003).

[11] Tom Fawcett, 'An introduction to {ROC} analysis', *Pattern Recognition Letters*, **27**(8), 861 – 874, (2006). {ROC} Analysis in Pattern Recognition.

[12] J. U. Garbas, T. Ruf, M. Unfried, and A. Dieckmann, 'Towards robust real-time valence recognition from facial expressions for market research applications', in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 570–575, (Sept 2013).

[13] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, 'Local binary patterns and its application to facial image analysis: A survey', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **41**(6), 765–781, (Nov 2011).

[14] Nils Ingemars. A feature based face tracker using extended Kalman filtering, 2007.

[15] R.E. Kalman, 'A new approach to linear filtering and prediction problems', *Journal of Basic Engineering*, **82**(1), 35–45, (March 1960).

[16] I. Kotsia, N. Nikolaidis, and I. Pitas, 'Fusion of geometrical and texture information for facial expression recognition', in *Image Processing, 2006 IEEE International Conference on*, pp. 2649–2652, (Oct 2006).

[17] I. Kotsia, S. Zafeiriou, N. Nikolaidis, and I. Pitas, 'Texture and shape information fusion for facial action unit recognition', in *Advances in Computer-Human Interaction, 2008 First International Conference on*, pp. 77–82, (Feb 2008).

[18] Sebastian Krappe, Michaela Benz, Thomas Wittenberg, Torsten Haferlach, and Christian Mnzenmayer. Automated classification of bone marrow cells in microscopic images for diagnosis of leukemia: a comparison of two classification schemes with respect to the segmentation quality, 2015.

[19] Sven E Krüger, Martin Schafföner, Marcel Katz, Edin Andelic, and Andreas Wendemuth, 'Speech recognition with support vector machines in a hybrid system.', in *INTERSPEECH*, pp. 993–996. Citeseer, (2005).

[20] Christian Küblbeck and Andreas Ernst, 'Face detection and tracking in video sequences using the modified census transformation', *Image Vision Comput.*, **24**(6), 564–572, (June 2006).

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, 'The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression', in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 94–101, (June 2010).

[22] Zuheng Ming, A. Bugeau, J. L. Rouas, and T. Shochi, 'Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine', in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 06, pp. 1–6, (May 2015).

[23] Ahmad Poursaberi, Hossein Ahmadi Noubari, Marina Gavrilova, and Svetlana N. Yanushkevich, 'Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification', *EURASIP Journal on Image and Video Processing*, **2012**(1), 1–13, (2012).

[24] Ognjen Rudovic, Mihalis A Nicolaou, Vladimir Pavlovic, R Walecki, O Rudovic, V Pavlovic, M Pantic, S Eleftheriadis, O Rudovic, M Pantic, et al., 'Machine learning methods for social signal processing', *Social Signal Processing*, **30**, 469–484, (2014).

[25] Tobias Ruf, Andreas Ernst, and Christian Küblbeck, 'Face detection with the sophisticated high-speed object recognition engine (shore)', in *Microelectronic Systems*, 243–252, Springer, (2011).

[26] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn, 'Deformable model fitting by regularized landmark mean-shift', *Int. J. Comput. Vision*, **91**(2), 200–215, (January 2011).

[27] E. Sariyanidi, H. Gunes, and A. Cavallaro, 'Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **37**(6), 1113–1133, (June 2015).

[28] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, 'Facial action recognition combining heterogeneous features via multikernel learning', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **42**(4), 993–1005, (Aug 2012).

[29] M. F. Valstar and M. Pantic, 'Fully automatic recognition of the temporal phases of facial actions', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **42**(1), 28–43, (Feb 2012).

[30] Michel F. Valstar and Maja Pantic, *Human–Computer Interaction: IEEE International Workshop, HCI 2007 Rio de Janeiro, Brazil, October 20, 2007 Proceedings*, chapter Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics, 118–127, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[31] Paul Viola and MichaelJ. Jones, 'Robust real-time face detection', *International Journal of Computer Vision*, **57**(2), 137–154, (2004).

[32] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng, 'Probability estimates for multi-class classification by pairwise coupling', *J. Mach. Learn. Res.*, **5**, 975–1005, (December 2004).

[33] Jizheng Yi, Xia Mao, Lijiang Chen, Yuli Xue, and A. Compare, 'Facial expression recognition considering individual differences in facial structure and texture', *IET Computer Vision*, **8**(5), 429–440, (October 2014).

[34] Hui Yu and Honghai Liu. Combining appearance and geometric features for facial expression recognition, 2015.

[35] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, 'A survey of affect recognition methods: Audio, visual, and spontaneous expressions', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**(1), 39–58, (Jan 2009).

[36] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran, *Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, November 13-17, 2011, Proceedings, Part III*, chapter Geometry vs. Appearance for Discriminating between Posed and Spontaneous Emotions, 431–440, Springer, Heidelberg, 2011.

[37] Yongmian Zhang and Qiang Ji, 'Active and dynamic information fusion for facial expression understanding from image sequences', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(5), 699–714, (May 2005).