Dynamic facial expression recognition by joint static and multi-time gap transition classification

Arnaud Dapogny, Kevin Bailly, Séverine Dubuisson

Dynamic facial expression recognition by joint static and multi-time gap transition classification

Arnaud Dapogny[1] and Kevin Bailly[1] and Séverine Dubuisson[1]
[1] CNRS, UMR 7222, ISIR, F-75005, Paris, France

***Abstract*— Automatic facial expression classification is a challenging problem for developing intelligent human-computer interaction systems. In order to take into account the expression dynamics, existing works usually make the assumption that a specific facial expression is displayed with a pre-segmented evolution, *i.e.* starting from neutral and finishing on an apex frame. In this paper, we propose a method to train a transition classifier from pairs of images. This transition classifier is applied at multiple time gaps and the output probabilities are fused along with a static estimation. We eventually show that our approach yields state-of-the-art accuracy on popular datasets without exploiting any such prior on the segmentation of the expression, thus effectively bridging the gap towards facial expression recognition in unconstrained environments.**

I. Introduction

Affective computing and more specifically automatic analysis of facial expressions from video sequences is a key to human-machine interactions. Applications are multiple, ranging from computer graphics animation to social monitoring or e-learning.

However, performing static facial expression classification from separate images is a challenging task as we observe large variability of lighting conditions as well as in subjects' morphology. Moreover, important pose variation may occasionally result in facial occlusions. Such difficulties can be alleviated by applying normalisation w.r.t. a neutral face representation of a specific subject [1], but expressions such as sadness or anger usually involve subtle face deformations that are hardly noticeable using only static information.

These subtle expressions are generally better understood by analysing the evolution of the facial deformation over time. In order to do so, the so-called dynamic approaches work under the assumption that specific facial expression video sequences are pre-segmented and that the expression evolves from a neutral face representation to the peak of a particular facial expression [2], [3], [4]. However, such prior does not hold in more realistic scenarios, where one has to jointly address the problem of temporal segmentation and expression recognition.

In this work, we propose a new method for facial expression recognition from video sequences, that uses static and transition classifications. Both static and transition classifiers are built upon geometric and appearance features in order to robustly estimate facial expression probabilities. Transition classification probabilities are estimated from multiple time gaps in the sequence then fused along with static probabilities in order to provide an expression prediction. The contributions of this work are thus three-fold:

1) A method for constructing a transition classifier from a set of image pairs that can be sampled from popular facial expression datasets.
2) A fusion model that efficiently combines static and transition information from different time gaps to recognize facial expression using both static and dynamic information.
3) A complete system that performs real-time facial expression classification from video sequences without the need of any explicit segmentation of the facial expression sequence.

To the best of our knowledge, this is the first time that such an approach is proposed that uses transition information between video frames from different time gaps to provide expression classification without requiring any prior segmentation or preprocessing of the sequences.

The rest of the paper is organized as follows: in Section II we review some existing works covering the field. In Section III we discuss the proposed method for building static and transition classifiers and combining these to apply multi-time gaps expression recognition on a video sequence. In Section IV we report evaluation results for two popular benchmarks. Finally, in Section V we present our conclusions and future works.

II. Related Work

Methods that exist in the literature for automatic facial expression recognition generally belong to either static or dynamic classification systems, as highlighted in [23].

On the one hand, static methods aim to recognize the expressions on each image in a separate way, without considering the relationship between the video frames. For instance, Shan *et al.* [5] introduce a static facial expression recognition system that consists of SVM classifiers built upon LBP features. Senechal *et al.* [6] propose to combine AAM coefficients and LGBP histograms. Zhi *et al.* [7] use a graph-preserving sparse NMF in order to develop an image representation that is well-suited for the classification task. Such approaches offer the advantage not to require any temporal pre-segmentation of the video sequences. However, they generally suffer from the inherent limitation of not using dynamic information to disentangle subtle emotions such as anger and sadness.

This work has been partially supported by the French National Agency (ANR) in the frame of its Technological Research CONTINT program (JEMImE, project number ANR-13-CORD-0004).

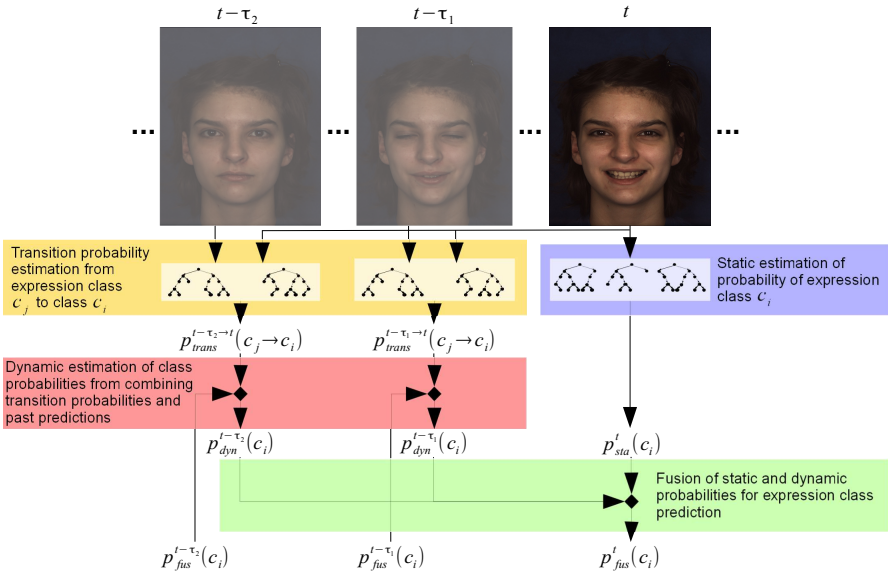

Fig. 1. Flowchart of the proposed facial expression recognition framework. For each new frame, facial feature points are extracted using the algorithm from [16]. Geometric and appearance features for both static and transition classifiers are computed for the current frame and between this current frame and the previous ones respectively. Static and transition classification probabilities sampled at multiple time gaps are then fused to provide the final expression prediction.

On the other hand, dynamic methods either use spatio-temporal features or temporal classifiers. Spatio-temporal features capture the motion of feature points or the evolution of a texture descriptor over time. For example, Zhao *et al.* extend LBP features [5] to spatio-temporal volumes [8]. Yang *et al.* [2] learn haar-like features from video cuboids using a boosting algorithm. Shojaeilangari *et al.* [9] compute histograms of phase and orientation as local video descriptors used for classification. Such approaches generally sample image features from a rigid time window which supposes a constant dynamic of the expression. Temporal classifiers describe the evolution of the expression from segments of a sequence. Jeni *et al.* [3], as well as Rudovic *et al.* [4], use SVM and ordinal regression from geometric features respectively. Lorincz *et al.* [10] compute dynamic programming kernels from facial feature point motion. Wang *et al.* [11] propose an interval-based temporal model to capture temporal relations between primitive facial events. However, these methods work under the assumption that the facial expression sequences are pre-segmented. Furthermore, sequence-level approaches have fewer training and testing samples as compared to frame-level methods.

Recently, some methods have been introduced that investigate temporal transitions for facial Action Unit (AU) recognition. Ding *et al.* [12] introduce a three-levels cascade of tasks where dynamic information is used to learn temporal segments of AU occurrence on a fixed-size window. Transition information is thus used for the purpose of refining the AU occurence boundaries. Khademi *et al.* perform relative AU recognition [13] using classifiers trained upon transition features from neighbouring frames. However, they do not consider static information and use the transitions as a reference in order to estimate AU intensity for current frame. Overall those two methods specifically aim to estimate AU activation over time. As compared to AUs which are more local in space and time, expressions describe a more global representation. As such, they generally last longer than AUs and involve larger facial deformations. For those reasons, expression recognition is often performed as a multiclass problem as compared to AUs which are evaluated separately.

III. Static and Multi-Time Gaps Transition Joint Classification

A. Overview

In this section we describe our static and multi-time gaps transition joint classification method. The framework involves the following steps: first, facial feature points are extracted using the algorithm from [16], then we perform a **static probability estimation (Section III-B)** upon separate video frames along with a **transition probability estimation (Section III-C)** between image pairs sampled at multiple time gaps. Those transition probabilities are then combined with past predictions to give rise to a **dynamic estimation of expression class probabilities (Section III-D)**. Finally, the system performs a **fusion of static and dynamic probabilities from multiple time-gaps (Section III-E)**.

B. Static probability estimation

Static facial expression classification is performed as a 7-class problem (the six universal expressions, plus the neutral one) using the Random Forest (RF) framework. RFs [14] are classifiers that are naturally suited for multiclass classification tasks. Their performance is on par with the most popular machine learning methods such as SVM or Neural Networks [14]. Furthermore, the RF framework provides the use of parallel implementation for the training step, as well

as an easily computable error estimate for an efficient testing procedure [15].

More specifically, a set of M decision trees is built upon the training dataset by a classic greedy procedure for RF classification:

- for m from 1 to M:
  1) Generate a bootstrap by randomly sampling the dataset.
  2) If the class repartition of the different classes among the bootstrap subset is imbalanced, a simple downsampling procedure is applied: samples of the majoritary class are randomly drawn out of the bootstrap until an acceptable imbalance level is reached.
  3) If the data at current node (initially at the tree root) is homogeneous with class c_i, then a terminal node is set, and terminal probabilities $p_m(c_i)$ of facial expression are set to 1 for c_i and 0 for other classes.
  4) If the classes are not homogeneous, we randomly generate a set of F_s binary features.
  5) For each binary feature, the induced data split is simulated and the corresponding entropy is computed. The split at current node is then set according to the feature that minimizes the entropy criterion.
  6) Go back to Step 3 for recursive application of the procedure on the induced subtrees.

Finally, static estimation of the probabilities of expression class c_i are computed as the average probability among the M trees of the forest, as shown in Equation (1).

$$p_{sta}(c_i) = \frac{1}{M} \cdot \sum_{m=1}^{M} p_m(c_i) \tag{1}$$

It is worth noting that, according to [14], the strength of the final classifier is a trade-of between the strength of each individual tree as well as the independence among the trees. Hence the downsampling step 2 allows the construction of individually weaker but more independent trees that, in our experiments, leads to better classification rates that we attribute to a better exploration of the feature space.

At Step 4, we generate $F_s = F_s^1 + F_s^2 + F_s^3$ features with:

1) F_s^1 distance features between two fiducial points provided by the SDM tracker [16]. Those distances are normalized w.r.t. inter-ocular distance to generate features that are invariant from face scaling.
2) F_s^2 angle features between three fiducial points.
3) F_s^3 HoG features sampled at any position of the face that is localized by its barycentric coordinates w.r.t. three fiducial points.

Each of the F_s features is associated with a threshold randomly sampled from a uniform distribution to produce binary features. For HoG features, we associate a threshold to one bin of the descriptor.

C. Transition probability estimation

In this subsection, we review how we use the dynamic information from pairs of images to classify transitions between two facial expressions. To this end, we use the same restricted transition model as in [17], where each facial expression state can only be reached by itself or *via* the neutral state (Figure 2).

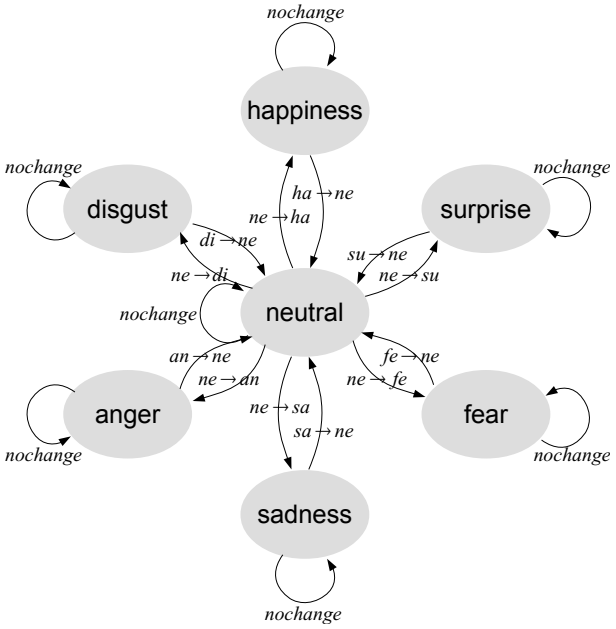

Fig. 2. Restricted transition model

In this model, we merge all the transitions that stay in a specific state into a $nochange$ class because from the perspective of classifying transitions it seems hard to determine in which static class the system lies when the two images are very close to each other. Hence, transition recognition is solved as a 13-label classification problem: $ne \rightarrow exp$ the transition class between neutral and a facial expression class among the six universal expression classes, $exp \rightarrow ne$ the transition class between an expression class and the neutral state, as well as the $nochange$ class.

Note that this model could be easily adapted to an ergodic expression transition model. However, as it is, it allows us to easily construct an image pair dataset by sampling specific expression sequences from the datasets, to which we associate the corresponding transition class. for instance, a record session with static expression class *happiness* provides examples of transition classes $nochange$, $ne \rightarrow ha$ and $ha \rightarrow ne$.

As in Section III-B, transition classification is performed using RF classifiers from image pairs. However, as will be discussed in the following subsections, the transition classifier will be applied several times for each new frame of a video stream. Thus, in order to ensure a very fast evaluation, we use a restricted set of $F_d = F_d^1 + F_d^2$ features generated from only two feature templates for splitting tree nodes:

1) F_d^1 differences of normalized point distances between the two images.
2) F_d^2 intensity differences of pixels localized by barycentric coordinates w.r.t. three fiducial points.

Benchmarks were conducted that shows that using this different set of feature templates for transition modeling leads to very similar video classification results, although using simple pixel comparisons instead of HoG features leads to a much lower CPU load.

In what follows, we refer to $p_{trans}^{t-\tau \to t}(c_i \to c_j)$ as the transition probability from expression class c_i to class c_j between images sampled at time $t-\tau$ and t of a sequence respectively. As it was done for static classification, we compute these probabilities by averaging the output probabilities among all the trees of the forest.

D. Dynamic estimation of class probabilities

We denote $p_{dyn}^{t-\tau}(c_i)$ the dynamic estimation of facial expression class c_i probability at time t given only the observations at times $t-\tau$ and t. This probability is computed as the most probable path between the two frames using past predictions $p_{fus}^{t-\tau}$ at time $t-\tau$ as well as the output of the transition classifier $p_{trans}^{t-\tau \to t}$ applied for images at time $t-\tau$ and t. (Equation (2)).

$$p_{dyn}^{t-\tau}(c_i) = max_{c_j}\{p_{trans}^{t-\tau \to t}(c_j \to c_i).p_{fus}^{t-\tau}(c_j)\} \quad (2)$$

In practice, as we use the simplified transition model from Figure 2, for each expression class we only need to compare the probabilities of a transition of type $ne \to exp$ and $nochange$. For the $neutral$ class, we have to examine the $nochange$ case, as well as any transition coming from an expression class.

E. Fusion of static and dynamic probabilities from multiple time-gaps

Static (Section III-B) and dynamic estimations (Section III-D) at multiple time gaps $\tau_0, \tau_1, ..., \tau_{k-1}$ are fused to give rise to the final prediction of class probability $p_{fus}^t(c_i)$ of expression class c_i at time t. In what follows, we refer to the number of time gaps k as the model order. A 0 order model is thus merely a static classifier that is applied separately on each video frame.

The prediction of class probability $p_{fus}^t(c_i)$ at time t can thus be computed from static and dynamic estimations in two different ways:

1) By averaging the probabilities $p_{dyn}^{t-\tau}(c_i)$ obtained for every time gap τ along with static estimation of class probability $p_{sta}^t(c_i)$ (Equation (3))

$$p_{fus}^t(c_i) = p_{sta}^t(c_i).(1 + \lambda \sum_{\tau} p_{dyn}^{t-\tau}(c_i)) \quad (3)$$

   where λ is a cross-validated constant that regulates the influence of dynamic information. The predictions p_{fus}^t are then renormalized to unit $\mathcal{L}_1$ norm.
2) By having static estimation and each dynamic estimation "vote" for the maximum probability class. Probabilities p_{fus}^t are then computed by averaging these votes.

A comparison between the average and the vote approaches can be found in Section IV-A.

IV. Evaluation

In order to validate our approach, we report accuracy on two broadly used facial expression recognition datasets for models of order 0 (static), 1 (with time gap $T = \{6\}$, *i.e.* we only look at one time gap six frames before the current one), 2 ($T = \{6, 12\}$), 3 ($T = \{6, 12, 24\}$) and 4 ($T = \{6, 12, 24, 48\}$), showing that dynamic estimations provided by our method allows to substantially enhance the recognition accuracy. Also note that the same benchmarks were conducted with different time gap values (e.g. $T = \{5, 10, 20, ...\}$, leading to similar conclusions.

The Extended Cohn-Kanade (CK+) database [18] includes 118 subjects for each of which from 1 to 11 expression recording sessions are associated. Some of these sessions do not belong to a specific expression class or are annotated with nonbasic label *contempt* and will thus be discarded for further processing. Videos presents acted expressions only and contains very few head pose variations. Table I shows the performance comparison on the CK+ database. For comparison purposes, results are reported for static/dynamic classification on the last frame of the videos.

The BU-4DFE database [19] contains 101 subject, each one displaying the 6 facial expressions. As in CK+, these records include few facial pose variations and the expressions are still acted, but the variability for facial expression is larger. Furthermore, video sequences in BU-4DFE do not necessarily end on a peak frame of an expression. Thus, we aknowledge correct classification for a video if the greatest sum of predicted probabilities greatest among the six expression correspond to the ground truth label. Table II displays the classification accuracies for the different expression classes obtained for BU-4DFE database.

Both static and transition classifiers are evaluated using the Out-Of-Bag (OOB) error estimate [14]. More specifically, during training bootstraps for static and transition classifiers are generated at the subject level so that, for the test, each tree is used only for subjects that do not belong to the corresponding bootstrap. OOB error, albeit being an unbiased estimate of the true generalization error, is easier to compute than traditional Leave-One-Subject-Out (LOSO) or k-fold cross-evaluation procedures. Also, it has been shown to be generally more pessimistic than traditional error estimates [20], further empathizing the quality of the proposed approach.

Last but not least, we also conducted a CPU load benchmark so as to evaluate the real-time capacities for our system.

A. Video sequence classification

Figure 3 shows the evolution of classification accuracies versus the model order respectively for CK+ and BU-4DFE databases. Table I shows the performances obtained using the fourth order model, as well as a comparison with

static classification for CK+ database. Table II displays the classification accuracies for the different expression classes obtained for the BU-4DFE database.

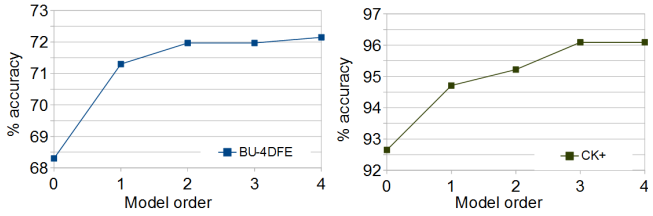

Fig. 3. Average accuracy vs model order for BU-4DFE and CK+ datasets

Expression	Order 0 (%)	Order 4 (avg)(%)	Order 4 (vote)(%)
Happiness	97.1	**98.6**	**98.6**
Anger	86.7	**95.6**	93.3
Sadness	92.9	92.9	92.9
Fear	88.0	**92.0**	88.0
Disgust	94.9	**100.0**	98.3
Surprise	96.4	**97.6**	**97.6**
Average	92.7	**96.1**	94.8

TABLE I

CLASSIFICATION ACCURACIES (%) FOR FOURTH ORDER MODEL ON THE CK+ DATASET

Expression	Order 0 (%)	Order 4 (avg)(%)	Order 4 (vote)(%)
Happiness	85.0	85.0	**86.0**
Anger	57.0	70.0	**75.0**
Sadness	**81.0**	80.00	79.0
Fear	34.0	36.0	**52.0**
Disgust	69.0	73.0	**74.0**
Surprise	82.8	**88.9**	**88.9**
Average	68.3	72.2	**75.8**

TABLE II

CLASSIFICATION ACCURACIES (%) FOR FOURTH ORDER MODEL ON THE BU-4DFE DATASET AND COMPARISON BETWEEN AVERAGE AND VOTE FUSION STRATEGIES

Work	#Subjects	Dynamic	Protocol	Acc(%)
LBP / SVM [5]	96	N	10-fold	92,6
LBP-TOP / SVM [8]	97	Y	10-fold	96,1
LPLO [9]	118	Y	10-fold	94,6
LSH-CORF [4]	98	Y	10-fold	86,8
Shape [3]	118	Y	LOSO	96
Ours, order 4 (vote)	118	N	OOB	94,8
Ours, order 4 (avg)	118	Y	OOB	96,1

TABLE III

CLASSIFICATION ACCURACY COMPARISON WITH STATE-OF-THE-ART METHODS ON CK+ DATASET

Classically, *anger* and *fear* are the facial expression classes for which the static classification accuracies are the lowest on both databases. Respective accuracies for these classes are 86.7% and 88.0% for CK+, and 70.0% and 36.0% for BU-4DFE. This is mainly due to the fact that the facial deformation can be quite subtle, and sometimes hard to

Work	#Subjects	Preprocessing	Acc(%)
BoMW [21]	101	None	63,9
Dynamic 2D [22]	101	None	67,0
LSH-CORF [4]	30	segmentation	77,1
Shape [3]	101	segmentation + neutral normalization	70,5
Shape [3]	101	segmentation + personal mean normalization	78,2
Ours, order 4 (avg)	101	None	72,2
Ours, order 4 (vote)	101	None	75,8

TABLE IV

CLASSIFICATION ACCURACY COMPARISON WITH STATE-OF-THE-ART METHODS ON BU-4DFE DATASET

analyse even for a human eye. Furthermore, the *fear* class is particularly difficult to act, and the facial expression sometimes look very similar to *surprise* or *happy*.

The addition of transition information allows to substantially increase the performance on those expression classes, as well as to help discriminate between *happiness*, *disgust* and *surprise*. Overall recognition percentages rises from 92.7% to **96.1%** for CK+ database, and from 68.3% to **72.2%** for BU-4DFE database for the average fusion method. Overall accuracy rises up to **75.8%** on BU-4DFE database when using the voting fusion scheme. Interestingly enough, the voting approach seems to perform slightly worse on CK+, which we believe is due to the shortness of the sequences in this database.

However, taking into account time steps beyond $\tau = 24$ does not typically add much to the classification results. This is specially true on the CK+ database where the recording sessions are rather short. Moreover, in less constrained benchmarks, it could very well decrease the accuracy as larger variations in face pose are to be observed for such large time steps, as well as transitions between different facial expression classes that may not be expected in the transition model shown in Figure 2.

Accuracy seems to be on par with state-of-the-art methods, as shown in Tables III and IV, although comparisons are to be put into perspective as the selected subsets and evaluation protocols are not exactly the same between the different approaches. Our static classifier gives similar results as LBP features with SVM classification [5], and classification accuracy for the fourth order model is also quite similar to LBP-TOP with SVM [8] as well as SVM from shape information [4]. As for BU-4DFE dataset, Our approach clearly outperforms other dynamic approaches that does not apply manual pre-segmentation of the sequences [21], [22]. Performances are also nearly equivalent to those reported by Rudovic *et al.* [4] and Jeni at al [3], although these methods imply manual truncation of the videos and/or normalization w.r.t. the first frame of the sequence [3], from which the provided accuracy values seems to be highly dependant.

B. Real-time capacities

Figure 4 shows the evolution of classification accuracy for the static and transition classifiers versus the number of trees, for both CK+ and BU-4DFE datasets.

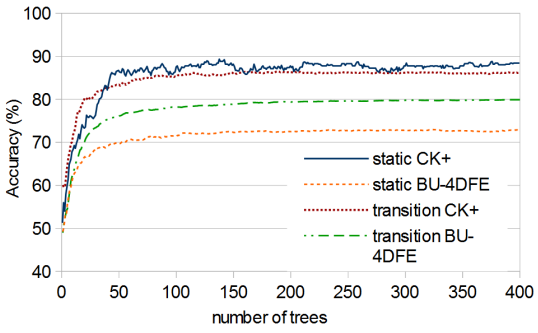

Fig. 4. Static and transition classification accuracy vs number of trees

It can be seen that overall accuracy reaches 95% of the maximum classification rates with only 50 trees. Using such setting, the proposed system runs at more 60 than fps. This evaluation includes facial alignment, feature extraction and expression classification steps. Facial alignment is the bottleneck of the system as the feature extraction and classification pipeline can be performed at more than 500 fps, although there is still room for code optimization and efficient parallelization. This benchmark was conducted on a Intel Core I7-4770 CPU with 32 Go RAM using a multithreaded C++/OpenCV environment.

V. Conclusion and Future Work

In this paper we have introduced a real-time, automatic facial expression recognition system that combines static and transition information from different time gaps. We propose a framework to train a transition classifier upon image pairs that can be extracted from popular datasets. The classifiers employ basic geometric and appearance features that provide robustness against variation in pose or in lighting conditions. The algorithm works in real time on a standard computer even for higher order models, without extensive code optimization.

Future work will involve applying our framework to real-case scenarios such as spontaneous facial expression datasets. For that matter we might have to adapt our transition model to efficiently take into account transitions from one expression state to another one, which may happen in such unconstrained scenarios. Furthermore, it will also be necessary to incorporate pose handling into the facial expression learning framework.

References

[1] Thibaud Senechal, Kevin Bailly, and Lionel Prevost. Automatic facial action detection using histogram variation between emotional states. In *ICPR*, pages 3752–3755, 2010.

[2] Peng Yang, Qingshan Liu, and Dimitris N Metaxas. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30(2):132–139, 2009.

[3] László A Jeni, Daniel Takacs, and Andras Lorincz. High quality facial expression recognition in video streams using shape related information only. In *ICCV Workshops*, pages 2168–2174, 2011.

[4] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In CVPR, pages 2634–2641, 2012.

[5] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Robust facial expression recognition using local binary patterns. In *ICIP*, volume 2, pages 370–373, 2005.

[6] Thibaud Senechal, Vincent Rapp, Hanan Salam, Renaud Seguier, Kevin Bailly, and Lionel Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):993–1005, 2012.

[7] Ruicong Zhi, Markus Flierl, Qiuqi Ruan, and Bastiaan Kleijn. Facial expression recognition based on graph-preserving sparse non-negative matrix factorization. In *ICIP*, pages 3293–3296, 2009.

[8] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[9] Seyedehsamaneh Shojaeilangari, Wei-Yun Yau, Jun Li, and Eam-Khwang Teoh. Multi-scale analysis of local phase and local orientation for dynamic facial expression recognition. *Journal ISSN*, 1(1), 2014.

[10] Andras Lorincz, Laszlo Attila Jeni, Zoltan Szabo, Jeffrey F Cohn, and Takeo Kanade. Emotional expression classification using time-series kernels. In *CVPR Workshops*, pages 889–895, 2013.

[11] Ziheng Wang, Shangfei Wang, and Qiang Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *CVPR*, pages 3422–3429, 2013.

[12] Xiaoyu Ding, Wen-Sheng Chu, Fernando De la Torre, Jeffery F. Cohn, and Qiao Wang. Facial action unit detection by cascade of tasks. In *ICCV*, pages 2400-2407, 2013.

[13] Mahmoud Khademi and Louis-Philippe Morency. Relative facial action unit detection. In *WACV*, pages 1090–1095, 2014.

[14] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[15] Leo Breiman. Out-of-bag estimation. Technical report, Statistics Department, University of California Berkeley, Berkeley CA 94708, 1996.

[16] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[17] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1):160–187, 2003.

[18] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010.

[19] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *FGR*, pages 211–216, 2006.

[20] Bylander Tom. Estimating generalization error on two-class datasets using out-of-bag estimates In *Machine Learning*, volume 48, pages 287–297 2002.

[21] Xu Liefei and Philippos Mordohai. Automatic Facial Expression Recognition using Bags of Motion Words. In *BMVC*, volume 10, 2010.

[22] Yi Sun and Lijun Yin. Facial expression recognition based on 3D dynamic range model sequences. In *ECCV*, pages 58–71, 2008.

[23] Zhihong Zeng, Maja Pantic, Glenn I. Roisman and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. In *Pattern Analysis and Machine Intelligence*, pages 39–58, 2009.