2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Facial Expression Recognition based on Static and Dynamic Approaches

Daniel Acevedo*‡, Pablo Negri†‡, María Elena Buemi* and Marta Mejail*
*Departamento de Computación, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires, Argentina.
Email: dacevedo@dc.uba.ar
†Universidad Argentina de la Empresa (UADE),
Lima 717, Buenos Aires, Argentina.
‡CONICET, Godoy Cruz 2290, Buenos Aires, Argentina.

***Abstract*—The identification of facial expressions with human emotions plays a key role in non-verbal human communication and has applications in several areas. In this work, we analyze two main approaches for expression recognition.**

One is a dynamic approach introducing a new simple descriptor based on the angles formed by the landmarks to capture the dynamic of the facial expression on a sequence. In this case the recognition is performed by a Conditional Random Field (CRF) classifier. An analysis of the most discriminative landmarks for this approach is presented.

The other is a static-based appearance method. In this approach, a binary-based descriptor, denominated Oriented Fast and Rotated BRIEF (ORB), is used on a single frame of a sequence of images to extract texture information, and classified with a Support Vector Machine.

We compare both methodologies, analyse their similarities and differences, and also propose simple combinations of both approaches to deal with their limitations.

I. Introduction

Facial expressions recognition has been a focus of interest for researchers for many years. Its connection to human emotions plays a key role in human communication and is approachable from several fields of study that lead to a vast range of applications, from psychology and marketing to human-robot interaction and pain assessment, among others. Another key role in human emotion recognition is Human Computer Interaction (HCI) since the expressiveness of human faces, which are usually linked to an emotional state, can be used to supersede other forms of non-verbal communication [1].

Psychology studies describe six basic emotions which are universally recognizable: happiness, sadness, surprise, fear, anger and disgust [2].

Several works have utilized local descriptors such as LBP since they have been successfully applied to model faces. Shan et al. [3] review a variety of techniques used for facial recognition and empirically evaluate LBP features with different machine learning methods applied on regular and low resolution images and videos. In [4] a new version of LBP is proposed and compared with 3DLBP for detection of 3D facial action units.

Sequence-based or dynamic classification is another methodology to solve facial expression recognition. The gesture dynamics can be captured by analyzing the landmarks locations and all relative deformations occurred from the Active Appearance Model (AAM) [5] based shape or neutral expression. Lucey et al. [6] and Jain et al. [7] employ Procrustes alignment [5] for this task. They obtain a 136 feature vector corresponding to the vertex displacements, which is the input for Support Vector Machine or Latent Conditional Random Fields classifiers respectively. In [8], they face the facial gestures recognition using dynamic textures approach: a temporal LBP feature is developed from three orthogonal planes corresponding to three consecutive frames. The classification task is performed by a SVM classifier with a second degree polynomial kernel function, and One-Against-One methodology.

In this work we aim at recognizing and classifying human emotions from image sequences. For that, we will make use of the Extended Cohn-Kanade AU-Coded Facial Expression Database (CK+), which has been designed for research in automatic facial image analysis and synthesis and for perceptual studies. In this database, several persons are recorded performing several facial gestures. For each of these gestures, the person starts from a neutral face and a sequence of images is obtained where the final image corresponds to the facial expression. The facial gestures in the database include seven basic emotions: anger, contempt, disgust, fear, happy, sadness and surprise. An extra neutral expression may be taken into account if the first image from the sequence is taken under consideration.

Two main approaches have been considered for emotion recognition.

In a dynamic approach, we introduce a new simple descriptor based on the angles formed by the landmarks placed on the images of faces that perform an expression. The dynamic of these angles is captured by means of a Conditional Random Field (CRF). We also perform an analysis for the selection of most convenient and discriminative landmarks.

A static approach is as well considered, where only the final image of the sequence (the one that ends with the facial expression) is used for training; a binary-based descriptor method is used on this single frame image and it is calculated on certain landmarks provided by the database. A Support Vector Machine is the classifier used for training and testing

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

in this static approach.

In this work we also propose simple combinations of the two aforementioned approaches so as to deal with the limitations that each approach has.

The paper is organized as follows: Section II introduces the proposed descriptor for the dynamic approach and the analysis for landmark selection. Section III describes the binary-based descriptor for the static approach. Section IV presents the experimental results for each approach and their combination. Finally, Section V presents conclusions and possible future work.

II. Dynamic Expression Recognition

A. Angular Descriptors

Fig. 1. The Figure shows an example of an angle generated by 3 different landmarks.

Expression recognition using temporal approaches usually tracks the changes of landmarks spatial positions within the images on the video sequence. This analysis, however, is sensitive to movement of the head while the expression is occurring.

Our descriptor, on the other hand, computes a measure independent of the spatial position of the landmarks on the image. The feature consists on the angle obtained by three landmarks. In this way, the approach is independent of the face pose, and the features only measure facial transformations by evaluating the changes on the angles between consecutive frames.

The total set of landmarks for each capture is 68. The number of angles is defined as $n!/((n-k)!k!)$ combinations of $n = 68$ items taken by $k = 3$ at a time. It results on more than 50,000 different angles, making impracticable their implementation. It is then necessary to select a subset of points to reduce this quantity. The criteria consists on identifying those landmarks that best capture the dynamics of the expressions through the sequence.

The methodology performs an exhaustive analysis on the 50,000 angles between the 68 landmarks on all the sequences.

Each angle is defined as $\alpha_{\ell_1,\ell_2,\ell_3}$, where ℓ_1 is the landmark at the central point, ℓ_2 and ℓ_3 are the extreme points of the angle. Fig. 1 presents examples on an angle generated by three landmarks.

The change of $\alpha_{\ell_1,\ell_2,\ell_3}$ in two captures separated by T frames is computed as $d_{\ell_1,\ell_2,\ell_3}(t) = \alpha_{\ell_1,\ell_2,\ell_3}(t) - \alpha_{\ell_1,\ell_2,\ell_3}(t-T)$. This difference measures the spatial evolution of the landmarks. Analyzing the sign and the magnitude of $d_{\ell_1,\ell_2,\ell_3}(t)$, the dynamic feature descriptor obtains three discrete values:

$$f_{\ell_1,\ell_2,\ell_3}(t) = \begin{cases} -1 & \text{if } d_{\ell_1,\ell_2,\ell_3}(t) < -\theta \\ 0 & \text{if } |d_{\ell_1,\ell_2,\ell_3}(t)| \leq \theta \\ +1 & \text{if } d_{\ell_1,\ell_2,\ell_3}(t) > \theta \end{cases} \quad (1)$$

where θ is a threshold that validates an angle change as significant. Its value depends on some parameters, such as the frame rate of the sequence, the size of the image, etc. In our experiments, setting $\theta = 5^o$ gives the best results.

Figure 2 shows a subset of landmarks and the evolution of two angles on two different expressions. On both examples, the angles are compared using a temporal interval of $T = 5$ frames. This value is closely related with the frame rate of the sequence, and must capture the dynamic of the gesture. If T has a low value, and the frame rate is high, the changes on the dynamic angles will not be noticeable.

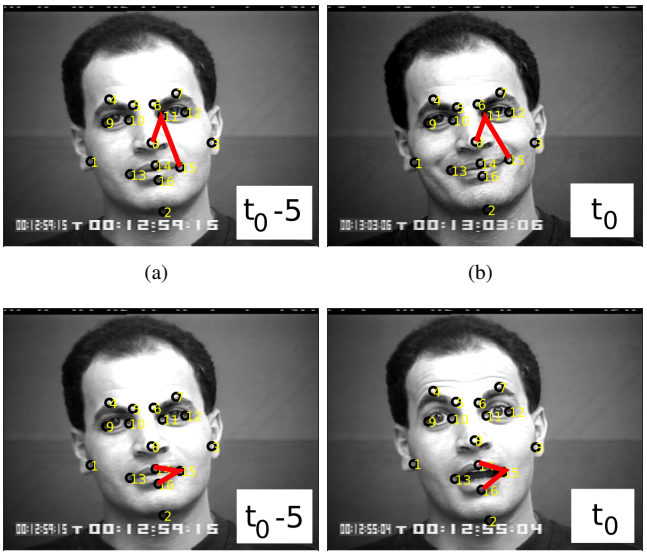

Fig. 2. First row shows two images of a happy gesture at frame $t_0 - 5$ (a) and at frame t_0 (b); the angle $\alpha_{\ell_{11},\ell_8,\ell_{15}}$ is centered at landmark ℓ_8, and in this case $f_{\ell_{11},\ell_8,\ell_{15}} = 1$ since $\alpha_{\ell_{11},\ell_8,\ell_{15}}(t_0) - \alpha_{\ell_{11},\ell_8,\ell_{15}}(t_0 - T) > \theta$. Second row of images (c) and (d) shows a similar example for a surprise expression and an angle formed by landmarks ℓ_{15}, ℓ_{14}, ℓ_{16}.

To select the subset from the 68 landmarks, we create seven accumulators, one for each expression, of length 68. When a dynamic feature $f_{\ell_1,\ell_2,\ell_3}(t)$ for a sequence of expression e is different from zero, the three landmarks ℓ_1, ℓ_2, and ℓ_3 receive one vote on the accumulator corresponding to e.

Fig. 3 shows the results, where the 20s most voted landmarks have darker color and red line. As can be seen, landmarks related to the mouth have the highest dynamics

along the expressions. Landmarks of eyebrows are also representative of *anger*, *disgust*, *fear*, and *sadness*. The nose have a high dynamic on *happy* and *disgust*. Finally, landmarks over the eyes are also representative of the *disgust* expression (for *surprise*, they also move a lot).

Based on the previous analysis, we choose a list of landmarks which have green color on the box at the bottom of Fig. 3. We also incorporate reference landmarks or *pivots* (marked in red) from those points which remain stable along the expression in order to improve robustness (some dynamics could be lost if we only choose moving landmarks).

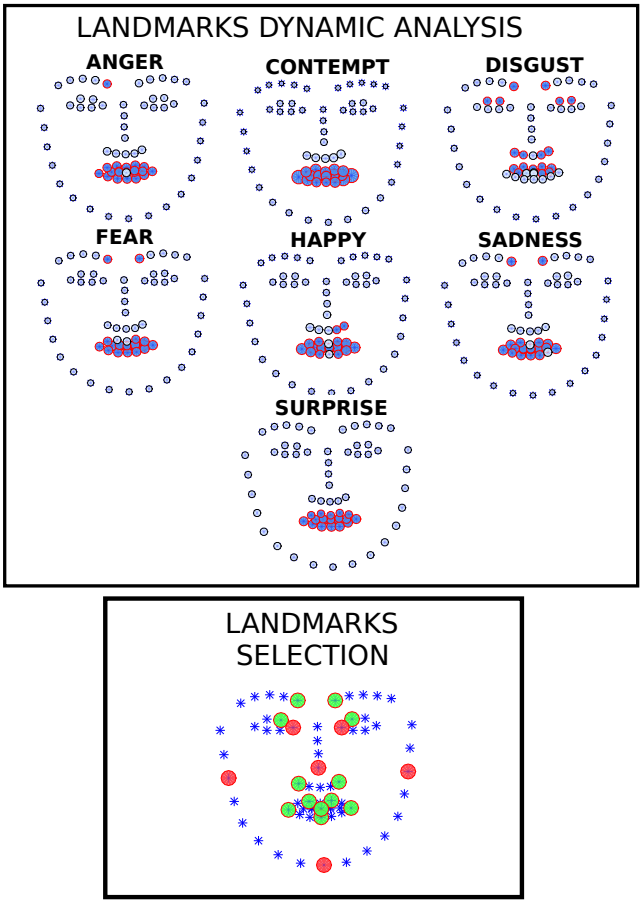

Fig. 3. Top: analysis of landmark dynamics. Darker dots indicate the most voted landmarks for each expression. Bottom: final selection of landmarks in red and green.

The set of dynamic features is defined using this subset of $n = 18$ landmarks. The feature vector of the frame at time t is $\mathbf{f}(t) = [f_{\ell_1,\ell_2,\ell_3}(t), f_{\ell_1,\ell_2,\ell_4}(t), \ldots, f_{\ell_{16},\ell_{17},\ell_{18}}(t)]$. The length of the vector is computed as $n!/((n-k)!k!)$ combinations of $n = 18$ items taken by $k = 3$ at a time, and resulting in 560 angles.

The choice of the central point ℓ_1 of each angle, between the three landmarks, is based on that point which maximizes the dynamics generated by the angle on all the sequences.

B. Dynamic Classification

Facial gestures, similarly to many other applications, have the particularity that the same gesture (for example, 'smile') can be performed at different speeds. Therefore, the number of frames of a video sequence capturing one person's gesture going from neutral to smiling, will differ from another person doing the same expression. In [9], Vainstein et al. faced the same kind of problem and compared a Support Vector Machine (SVM) and a Conditional Random Field (CRF) approaches to tackle tennis gestures recognition. SVM approach uses a framework called local features that defines a kernel function seeking feature similarities between test and train samples. Coefficients weighting the differences at the positions within the sequences are incorporated, in order to have reliable comparisons of the gesture. CRF, on the other hand, encodes by itself the temporal sequence of the descriptors on the sequence, and obtain the best results on the tennis dataset.

CRF is a statistical method that uses graphical models for predicting complex structures. In our work, we employ linear-chain CRF to model the sequential dependencies between the video frames. The gesture recognition can be regarded as a multivariate prediction problem, seeking to identify the sequence defined by $(\mathbf{y}, \mathbf{x})$, where $\mathbf{y} = \{y_0, \ldots, y_N\}$ are the tags of each frame, and $\mathbf{x} = \{\mathbf{x}_0, \ldots, \mathbf{x}_N\}$ are the corresponding feature vectors. The list of L labels $\{happy, sadness, \ldots, fear\}$ is also defined. Linear-chain CRF employs a conditional distribution $p(\mathbf{y}|\mathbf{x}, w)$ based on the log-linear model and the output variables given the observable feature vectors [10]:

$$p(\mathbf{y}|\mathbf{x}, w) = \frac{1}{Z(\mathbf{x}, w)} \prod_{k=1}^{L} \exp\left(\sum_{t=1}^{N} w_k \, \phi_k(y_{t-1}, y_t, \mathbf{x}_t)\right) \tag{2}$$

where ϕ_k are the feature functions associated to class k evaluating the compatibility between label y_t and feature vector $\mathbf{x}_t$, w_k are the class weights, and $Z(\mathbf{x}, w)$ is an instance normalization function that ensures the distribution p equals one.

III. Appearance Expression Recognition Using ORB

In this paper, we analyze ORB (Oriented Fast and Rotated BRIEF) [11] as a feature extraction method to estimate the facial expression in human faces. It is a fast robust local feature detector based on BRIEF (Binary Robust Independent Elementary Features) [12]. These descriptors are calculated from keypoints on the faces. For each keypoint a feature vector is extracted by ORB. The face descriptor is generated as the concatenation of the vectors obtained for each keypoint. In our work, landmarks provided by the database were used as keypoints.

The BRIEF descriptor [12] is a description of an image generated from a set of points of interest. For each keypoint k, a patch P_k of size $N \times N$ around k is considered. The feature

vector of P_k is constructed from a set of pairwise intensity comparisons. To do that, a test τ_k on P_k is defined as

$$\tau_k\left(P_k, a, b\right) = \begin{cases} 1, & \text{if } p_k\left(a\right) < p_k\left(b\right) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $p_k(a)$ and $p_k(b)$ are the intensities in a smoothed version of P_k at points a and b, respectively. The smoothed version of P_k is used to make the descriptor less sensitive to noise. The descriptor of P_k is defined as a vector of n binary tests:

$$f_n\left(P_k\right) = \sum_{i=1}^{n} 2^{i-1} \tau_k(P_k, a_i, b_i) \quad (4)$$

BRIEF does not have an elaborate sampling pattern to extract sample points in the region around the keypoint or an orientation compensation mechanism; thus, pairs can be chosen randomly at any point on the patch.

There are two main differences between ORB and BRIEF: 1) ORB uses an orientation compensation mechanism, making it rotation invariant. 2) ORB learns the optimal sampling pairs, whereas BRIEF uses randomly chosen sampling pairs.

The idea of ORB is to "steer" the BRIEF descriptor according to the orientation of keypoints. For each keypoint k, we have a set of pairs of points $S_k = \{(a_i, b_i)\}_{1 \leq i \leq n}$ where a_i and b_i are the two points in the patch P_k that define a binary test (see Eq. (3)). Using the patch orientation θ, calculated from the intensity centroid ("center of mass") of the patch [13], a steered version S_θ of S_k is built. To do that, points in S_k are rotated using the angle θ. The steered BRIEF descriptor is calculated as $f_n(P_k)$ using only the points in S_θ:

$$g_n\left(P_k, \theta\right) = f_n\left(P_k\right) \quad \text{for } (a_i, b_i) \in S_\theta \quad (5)$$

where P_k is the patch around k and f_n is the BRIEF descriptor, defined in Eq. (4).

The classification algorithm used to estimate the expression is SVM (Support Vector Machine) with a radial basis function kernel. The input of the classifier for each image is the set of descriptors extracted from all the landmarks provided by the database (see Section IV).

IV. Experiments and Results

We performed several experiments for evaluating the accuracy of emotion recognition and we analyzed and compared different methods to solve the problem. A leave-one-out subject cross validation methodology was used to assess the performance in the study: each frame of a sequence belonging to a subject in the database is classified (i.e., tested) and the rest of the subjects are used for training.

A. Database

In our experiments we utilized the extended Cohn-Kanade dataset (CK+) [6], which is one of the most widely used resources in the development of expression recognition systems. The CK+ database, contains sequences of frames that begin with a neutral expression (Ne) and end in one basic emotion: anger (An), contempt (Co), disgust (Di), fear (Fe), happiness (Ha), sadness (Sa) or surprise (Su). There are 118 subjects that perform from one to six expressions each, to a total of 327 sequences which are labeled as one of the seven aforementioned expressions (but no sequence is labeled as neutral on the database). The database also provides 68 landmarks for each frame of each sequence. Figure 4 shows an example of three frames of a sequence of the CK+ dataset; the target expression is surprise.

Fig. 4. Three frames of a sequence of CK+ dataset. It begins with a neutral expression (first frame) and proceeds to a target expression (last frame). In the example shown, the target expression is surprise.

B. Dynamic Angles + Conditional Random Fields

In this work, CRF is implemented using the software CRFSuite [14]. The training combines the BFGS method and the Orthant-Wise Quasi-Newton method. CRFSuite has the particularity that input feature vectors correspond to a list of tags or strings. Therefore, feature vector $\mathbf{f}(t)$ values on equation (1) are converted to a new feature vector $\mathbf{x}_t$ in the following way: 1 values get tag 'P' (positive), 0 values get 'Z', and -1 values get tag 'N' (negative).

To classify a video, the feature vectors $\mathbf{x}_t$ are evaluated sequentially, from frame $t = T$ to the last frame. Each $\mathbf{x}_t$ obtains from CRF a label indicating their corresponding gesture. The video is then classified with the gesture that obtained the greatest number of votes.

Table I presents the confusion matrix results corresponding to the dynamic angles and the CRF classification. The overall performance of the methodology is 80.5% and corresponds to the average of diagonal values of the confusion matrix.

	An	Co	Di	Fe	Ha	Sa	Su
An	**93.3**	0	4.4	0	2.2	0	0
Co	11.1	**55.6**	0	5.6	22.2	5.6	0
Di	6.8	0	**89.8**	0	3.4	0	0
Fe	8	0	0	**60**	16	12	4
Ha	1.4	0	0	0	**97.1**	0	1.4
Sa	14.3	0	3.6	3.6	0	**71.4**	7.1
Su	.2	1.2	0	1.2	0	0	**96.4**

TABLE I
CONFUSION MATRIX OF EXPRESSION RECOGNITION FOR THE DYNAMIC FEATURES AND CRF: ANGER (AN), CONTEMPT (CO), DISGUST (DI), FEAR (FE), HAPPINESS (HA), SADNESS (SA) AND SURPRISE (SU), EXCLUDING THE NEUTRAL EMOTION.

The poorest performance is found on the Contempt gesture (55%) which is mostly confused with Happy. This is because the Contempt gesture is performed by the actors with lips movement, similar to a smile, with the mouth closed. Fig. 5 shows two sequences exemplifying the contempt expression. It can be noticed that the final image of the sequence resembles a happy expression.

Fig. 5. Each row shows en example of a sequence from neutral expression (left) to the contempt expression (right).

We have observed that dynamic angles tend to lose information when the motion in the sequence is small. This usually happens on long sequences with subtle movements where dynamic angles are unable to capture the dynamics, turning out in missclassification. Results show that 32 sequences, within the 43 sequences wrongly classified, have at least 50% of frames considered as static (no motion is perceived).

C. Static Approach Using ORB + SVM

We performed two experiments to analyze the accuracy of ORB as a feature descriptor for face emotion recognition. The classification algorithm used to estimate the expression was SVM with a radial basis function kernel. We used OpenCV implementation for ORB and LIBSVM library for SVM algorithms.

Two different experiments were conducted on the database. In a first experiment, the seven categories of universal facial expressions are used as in the dynamic approach: anger, contempt, disgust, fear, happiness, sadness and surprise. In a second experiment, it was also considered the neutral emotion. For each image used in the experiments, the 68 landmarks were used as keypoints. For each landmark, $n = 256$ binary tests (see Eq. 4) were calculated in order to extract a feature vector.

For the basic expressions' training we only used the last frame of each sequence. For the tests, each frame of a sequences is classified, and the class of a sequence is determined according to the most numerous expression that is present in it. Results as a confusion matrix are shown in Table II.

	An	Co	Di	Fe	Ha	Sa	Su
An	**57.8**	2.2	11.1	2.2	0	26.7	0
Co	11.1	**66.7**	0	0	0	22.2	0
Di	1.7	3.4	**78**	3.4	3.4	10.2	0
Fe	8	4	16	**40**	12	8	12
Ha	0	5.8	2.9	1.4	**89.9**	0	0
Sa	0	7.1	14.3	0	0	**78.6**	0
Su	1.2	1.2	9.6	2.4	1.2	6	**78.3**

TABLE II
CONFUSION MATRIX OF EXPRESSION RECOGNITION FOR THE ORB FEATURES ON THE SEVEN BASIC EMOTIONS.

When the neutral expression is also considered, a first frame of a sequence from each subject in the database is added to the training set (labeled as neutral). Since the ground truth labels provided by the database for each sequence correspond to one of the seven performed basic expression (but not neutral), we will omit the frames classified as neutral when the most numerous expression in the sequence is determined. When all the frames are classified as neutral, only in this case, the whole sequence is classified as neutral and they are not omitted. A confusion matrix with an extra column considering the percentage of sequences classified as neutral is shown in Table III. This last column represents the percentage of sequences whose frames are all misclassified as neutral whose actual class is indicated by the corresponding row.

	An	Co	Di	Fe	Ha	Sa	Su	Ne
An	**80**	2.2	3.6	2.2	0	2.2	0	8.9
Co	5.6	**61.1**	0	0	5.6	5.6	0	22.2
Di	3.4	0	**91.5**	1.7	1.7	0	0	1.7
Fe	4	0	8	**60**	12	0	8	8
Ha	0	2.9	1.4	1.4	**94.2**	0	0	0
Sa	3.6	0	3.6	0	0	**71.4**	0	21.4
Su	0	1.2	0	2.4	0	1.2	**95.2**	0

TABLE III
CONFUSION MATRIX OF EXPRESSION RECOGNITION FOR THE ORB FEATURES AND CONSIDERING THE NEUTRAL EMOTION (NE).

The overall performance (the average of the confusion matrix diagonal percentage values) without considering the neutral emotion is 69.9%, and 79.1% when the neutral emotion is considered. We can see that the inclusion of the neutral expression improves classification for most of the expressions except for Contempt and Sadness. As with the dynamic approach, the most difficult expression to classify is Contempt.

D. Combining both approaches

In this work we have considered a simple way of merging both methodologies to tackle the shortcomings of each approach.

When using ORB, the main limitation of this method is that for some expressions that are performed in a subtle way, all the frames in the sequence were classified as neutral (these cases are contemplated in the last column of Table III). To solve this in a simple way, we decided to replace the ORB neutral prediction with the CRF prediction. Results for this combination methodology is presented in Table IV, achieving a classification rate of 85.9% on the average.

The dynamic angles method has also some deficiencies when the movements along the expression are minimal. This movement is represented by the number of angles that change between frames. When this number is below a threshold, we detected that CRF behaved erratically, and therefore its prediction was omitted. For these frames that no predictions were provided by CRF, we used the ones provided by ORB. Results are presented in Table V, with a classification rate of 82.1% on the average.

	An	Co	Di	Fe	Ha	Sa	Su
An	**88.9**	2.2	4.4	2.2	0	2.2	0
Co	5.6	**72.2**	0	5.6	11.1	5.6	0
Di	3.4	0	**93.2**	1.7	1.7	0	0
Fe	4	0	8	**68**	12	0	8
Ha	0	2.9	1.4	1.4	**94.2**	0	0
Sa	3.6	0	3.6	0	0	**89.3**	3.6
Su	0	1.2	0	2.4	0	1.2	**95.2**

TABLE IV
CONFUSION MATRIX OF EMOTION CLASSIFICATION FOR THE ORB FEATURES COMBINED WITH CRF CLASSIFICATION WHEN NEUTRAL SEQUENCES ARE DETECTED.

	An	Co	Di	Fe	Ha	Sa	Su
An	**95.6**	0	2.2	0	2.2	0	0
Co	11.1	**55.6**	0	5.6	22.2	5.6	0
Di	3.4	0	**94.9**	0	1.7	0	0
Fe	8	0	0	**64**	16	8	4
Ha	1.4	0	0	0	**97.1**	0	1.4
Sa	14.3	0	3.6	3.6	0	**71.4**	7.1
Su	1.2	1.2	0	1.2	0	0	**96.4**

TABLE V
CONFUSION MATRIX OF EMOTION CLASSIFICATION USING THE DYNAMIC ANGLES' FEATURES COMBINED WITH THE ORB CLASSIFICATION WHEN THE DYNAMIC OF ANGLES IS SMALL.

Considering the average of the diagonal percentage values in Tables IV and V, results from Table IV (85.9%) are better that those in Table V (82.1%). In case of Contempt, which is usually a hard expression to classify, the improvement is remarkable; so is the case with Sadness.

V. CONCLUSIONS

In this paper we have presented a simple descriptor for facial expression recognition based on angles whose dynamic is intended to be modeled by means of a Conditional Random Fields. As opposed to this intrinsic dynamic approach, we have made a comparison with an appearance-based method which is based on the ORB descriptor (not previously used in the expression recognition literature, as far as we are concerned). In the experiments presented, the CRF-based method has achieved slightly higher classification results (80.5% for CRF vs. 79.1% for ORB with the neutral expression).

It is important to remark the difference in size of both descriptors. The dynamic angles approach uses a subset of landmarks and amounts to a total of 560 angle values which are fed to the CRF for each frame. As well, each of these values can be represented with only three possible angle variations: positive, negative or zero (can be coded in 3 bits). An analysis on the angles that changed more frequently for each expression allowed us to reduce the number of angles to compute and only keep a minimum subset for all the expressions. On the other hand, we have used all the landmarks for the ORB descriptor resulting in 68×256 bits. Thus, with this aspect in consideration, we may conclude that the dynamic approach achieves comparable results to the static approach with a simpler and more compact representation.

The ORB-based approach has a clear advantage over the dynamic approach: it can be applied on a single frame and does not depend on the motion performed to reach the expression. On one hand, it can be used on the frames where the dynamic angles method fails to output a classification result because of the absence of movement in the sequence. On the other hand, this allows to include the neutral expression as an extra class to be classified. The CRF-based method falls short with the possibility to include this neutral expression since the lack of motion makes it produce no outcome or makes it behave erratically. Nonetheless, we found cases where the whole sequence was classified as neutral by ORB and the CRF still performed well.

Therefore, as shown on this work, by the combination of both methodologies we obtain significant improvements with respect to the implementation of ORB and dynamic angles separately (combined methods: see Table IV and V; ORB and CRF-based methods: see Table I y III, respectively).

There are several options for merging these two approaches which were not described in this work and could be suitable for their implementation in an on-the-fly recognition application. By means of a more detailed analysis on the dynamic of the expression (which can be gauged by the angle changes between consecutive frames) we foresee that there is room for improvement.

ACKNOWLEDGMENT

This work has been partially supported by UBA project UBACyT 20020130200290BA, and ACyT A15T14 of UADE.

REFERENCES

[1] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. on PAMI*, vol. 37, no. 6, pp. 1113–1133, June 2015.
[2] T. Dalgleish and M. J. Power, Eds., *Handbook of cognition and emotion*. New York: Wiley, 1999.
[3] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vision Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.
[4] G. Sandbach, S. Zafeiriou, and M. Pantic, "Local normal binary patterns for 3d facial action unit detection," in *ICIP*, 2012, pp. 1813–1816.
[5] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
[6] P. Lucey, J. Cohn, T. Kanade, J. Saragih, and M. I. Ambadar, Z., "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," *CVPRW*, pp. 94–101, 2010.
[7] S. Jain, C. Hu, and J. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *IEEE International Conference on Computer Vision Workshops*, Nov 2011, pp. 1642–1649.
[8] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. PAMI*, vol. 29, no. 6, pp. 915–928, 2007.
[9] J. Vainstein, J. Manera, P. Negri, C. Delrieux, and A. Maguitman, "Modeling video activity with dynamic phrases and its application to action recognition in tennis videos," in *CIARP*, vol. 8827, 2014.
[10] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," *ArXiv e-prints*, Nov. 2010, http://arxiv.org/abs/1011.4088.
[11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.
[12] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," *11th ECCV*, pp. 778–792, 2010.
[13] P. L. Rosin, "Measuring corner properties," *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 291 – 307, 1999.
[14] N. Okazaki, "CRFsuite: a fast implementation of Conditional Random Fields," 2007. [Online]. Available: www.chokkan.org/software/crfsuite/