

A Citation k-NN Approach for Facial Expression Recognition

Daniel Acevedo^{1,2}(✉), Pablo Negri³, María Elena Buemi¹,
Francisco Gómez Fernández^{1,2}, and Marta Mejail^{1,2}

¹ Facultad de Ciencias Exactas y Naturales, Departamento de Computación,
Universidad de Buenos Aires, Buenos Aires, Argentina

dacevedo@dc.uba.ar

² Instituto de Investigación en Ciencias de la Computación (ICC),
CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina

³ CONICET-Universidad Argentina de la Empresa (UADE),
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 propose a descriptor based on areas and angles of triangles formed by the landmarks from face images. We test this descriptors for facial expression recognition by means of an adaptation of the k-Nearest Neighbors classifier called Citation-kNN in which the training examples come in the form of sets of feature vectors. Comparisons with other state-of-the-art techniques on the CK+ dataset are shown. The descriptor remains robust and precise in the recognition of expressions.

1 Introduction

Facial expressions have been studied due to the great deal of applications that are required for interpreting human communication through facial gestures, such as automatic behavior recognition, human-computer interaction, pain assessment, health-care, surveillance, deceit detection and sign language recognition [1]. In Human Computer Interaction (HCI), it can be used to supersede other forms of non-verbal communication, since the expressiveness of faces are usually linked to an emotional state [2].

We aim at recognizing human emotions from the complete set of the Extended Cohn-Kanade AU-Coded Facial Expression Database (CK+) [3]. This database captures several persons performing facial expressions, including seven basic emotions: anger, contempt, disgust, fear, happy, sadness and surprise.

In this work, a geometric descriptor that captures facial expression evolution within a video sequence is introduced. It is based on the angles and areas formed by the landmarks placed on the images of faces that perform an expression. A similar descriptor was introduced in [4] and its dynamic was modeled with a Conditional Random Field (as opposed to our Multiple Instance Learning

approach presented in this work). Geometric descriptors were also proposed for other applications and different contexts. In [5] geometric descriptors are used for kinship verification. Angles from triangles formed by 76 landmarks are extracted from images as well as segment lengths and its ratios. In [6] the authors propose for age invariant face recognition the use of a set of triangles proportion ratios to estimate the degree of similarity among triangles. A thorough study of different methodologies can be found at [7].

We use the geometric descriptor (angles and areas) in a variant of a supervised machine learning framework in which the descriptors are grouped in a form of bags. By means of a distance between bags (Hausdorff distance), an adaptation of the k-nearest neighbor algorithm is used as a classifier, called Citation k-NN [8].

The paper is organized as follows: Sect. 2 introduces the proposed descriptor. Section 3 presents the adaptation of our geometric descriptor to be used by the Citation k-NN classifier. Section 4 presents the experimental results. Finally Sect. 5 presents conclusions.

2 Geometrical Descriptor

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

Our descriptor, computes a measure independent of the spatial position of the landmarks on the image. They are computed from the triangles obtained by three landmarks, inspired on the polygon mesh representation. 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 and areas on the respective triangles between consecutive frames. Three landmarks (ℓ_a, ℓ_b, ℓ_c) define a triangle Δ . The geometry of Δ can be described by its *area* and the *internal angles*. Thus, let be: θ_Δ , the angular descriptor where ℓ_a is the landmark at the central point, ℓ_b and ℓ_c are the extreme points of the angle, and α_Δ is the area of triangle Δ . Figure 1 shows a subset of landmarks and the evolution of an

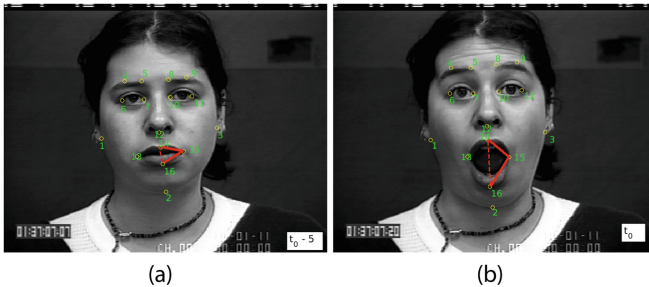


Fig. 1. Images of a surprise expression with landmarks numbered and the triangle $\Delta = (\ell_{14}, \ell_{15}, \ell_{16})$ displayed in red lines, (a) $t_0 - 5$ frame, centered at angle $\hat{\ell}_{15} = 45^\circ$, (b) t_0 frame, centered at angle $\hat{\ell}_{15} = 96^\circ$.

angle on a surprise expression using a temporal interval of 5 frames. Traditionally, the set of landmarks for each capture is composed of 68 points. We work with a subset of $n = 18$ landmarks that define 814 triangles describing the geometrical shape of the face and their dynamic.

3 Multiple Instance Learning and Citation-kNN

Multiple Instance Learning (MIL) is a variant of supervised machine learning in which the training examples come in the form of *bags*. A bag is a set of feature vectors, called *instances*. Each training bag has, as in traditional supervised learning, an assigned class label. The goal is to learn a model capable of predicting class labels for unseen query bags. Note how this setting considerably differs from classical single instance supervised learning, in which training data is organized as individual feature vectors, and a class label is available for each vector. A variety of approaches to the MIL task have been proposed over the years [8]. Every method relies on a specific assumption about the relationship between a bag label and the instances within that bag. Citation-kNN is a way of adapting the *k-Nearest Neighbors* (kNN) approach to the MIL task. It is based on the very loose assumption that bags that are similar according to a given distance measure are likely to have the same class label. In the single-instance variant of kNN, traditional distance functions such as L_p norms can be used to compare examples. In MIL though, proper distance functions between bags have to be used instead.

The chosen distance function in Citation-kNN is a modified version of the Hausdorff distance. Given two bags $A = \{a_1, \dots, a_M\}$ and $B = \{b_1, \dots, b_N\}$, and two integers S and L such that $1 \leq S \leq M$ and $1 \leq L \leq N$ the Hausdorff distance is defined as:

$$H_{SL}(A, B) = \max(h_S(A, B), h_L(B, A)) \quad (1)$$

The function $h_F(A, B)$ is known as the *directed Hausdorff distance*. Given bags A and B , and an integer F such that $1 \leq F \leq |A|$, it is defined as: $h_F(A, B) = F^{th} \min_{b \in B} \|a - b\|$, where $\|\cdot\|$ is a norm on the points of A and B and F^{th} is the F -th ranked distance in the set of distances (one for each element in A). In other words, for each point in A the distance to its nearest neighbor in B is computed, and the points of A are ranked according to the values of this distance. The F -th ranked value decides the distance $h_F(A, B)$.

The values of S and L are usually defined indirectly by specifying the fraction $0 \leq K \leq 1$ of the points to be considered. That is, given bags A and B , S and L are respectively set as: $S = \max(\lfloor |A|K \rfloor, 1)$ and $L = \max(\lfloor |B|K \rfloor, 1)$, where $\lfloor \cdot \rfloor$ is the floor function and $|\cdot|$ denotes cardinality.

The usual kNN method for selecting the label of a query bag based on the majority class of the closest neighbors may not lead to good classification results in the MIL setting. To overcome this problem, [9] incorporates the notion of *citation* into the voting mechanism used to decide the label of a query bag.

The proposed approach defines the R -nearest references of a bag A as the R nearest neighbors of A . Moreover, it defines the C -nearest citers of A to be the set that includes any given bag X if and only if A is one of the C nearest neighbors of X . For binary classification, the number of positive votes p is determined by summing the number of positive references R_p and positive citers C_p , i.e. $p = R_p + C_p$. Likewise, if the number of positive and negative references are R_n and C_n respectively, the number of negative votes is $n = R_n + C_n$. The query is predicted positive if $p > n$, and negative otherwise.

3.1 Modified Citation-kNN

The voting procedure described in Sect. 3 was thought for binary classification. However, a natural extension to the multiple class scenario can be devised for citation-kNN. Let L be the set of possible labels. For a given query bag X , assume $\{r_1, \dots, r_{n_r}\}$ and $\{c_1, \dots, c_{n_c}\}$ are the class labels of its n_r nearest references and its n_c nearest citers respectively. In our extension, the predicted label l^* for X is obtained by summing the votes of citers and references as:

$$l^* = \arg \max_{l \in L} \left(\sum_{i=1}^{n_c} \delta(l, c_i) + \sum_{j=1}^{n_r} \delta(l, r_j) \right) \quad (2)$$

where $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ otherwise.

The criteria presented in Eq. (2) assigns the same importance to the vote of any citer or reference. However, it seems reasonable to pay more attention to those neighbors that are closer to the query bag. This can be achieved by using weighted voting. Formally, let X be the query example, L be the set of possible labels, and assume $\{R_1, \dots, R_{n_r}\}$ and $\{C_1, \dots, C_{n_c}\}$ are the sets of the n_r nearest references and the n_c nearest citers of X respectively. Further, let $\{r_1, \dots, r_{n_r}\}$ and $\{c_1, \dots, c_{n_c}\}$ be the class labels of those nearest references and citers respectively. Then, the weighted voting approach for predicting label l^* for X is:

$$l^* = \arg \max_{l \in L} \left(\sum_{i=1}^{n_c} \alpha_i \delta(l, c_i) + \sum_{i=1}^{n_r} \beta_i \delta(l, r_i) \right) \quad (3)$$

where α_i and β_i are the weights for citer C_i and reference R_i respectively.

Clearly, the weight for a particular example should be a decreasing function of its distance to X . In this work we set $\alpha_i = \frac{1}{H(X, C_i)}$ and $\beta_i = \frac{1}{H(X, R_i)}$, where $H(\cdot)$ is the Hausdorff distance described in Eq. (1); we omit the S and L subscripts here for clarity.

3.2 Citation kNN on Bags of Time-Stamped Facial Expression Descriptors

Our MIL approach for facial expression recognition represents descriptor sequences using bags of time-stamped descriptors and classifies new bags using

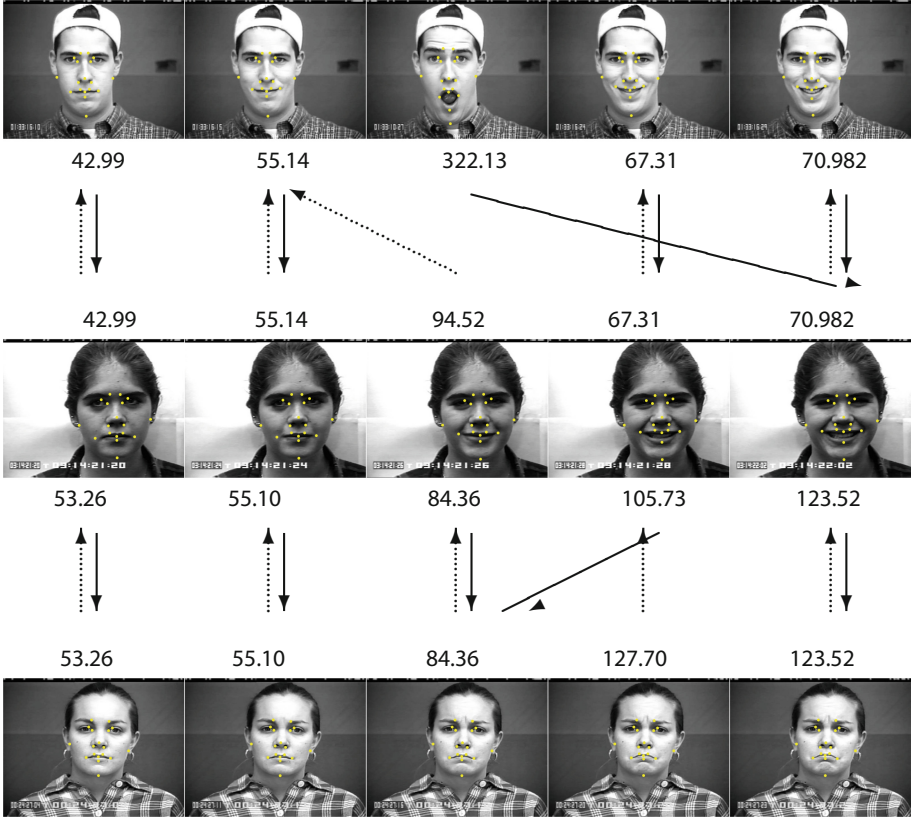


Fig. 2. Illustration of the Hausdorff distance between bags of face expression descriptors, corresponding to the *happy* (top and middle respectively) and *sadness* (bottom) emotions. Considering each pair of expressions (top-middle and middle-bottom), values at the beginning of each downward solid arrow describes the euclidean distance from one expression to the nearest expression below it, while values at the beginning of each upward dotted arrow describes the euclidean distance from one expression to the nearest above it.

Eq. (3). We now formalize the proposed representation and present a series of examples that aim at gaining further insight into the Hausdorff distance. For clarity purposes, the original sequences used in the examples were brought to a length of only 5 frames.

From the definition of areas and angles in Sect. 2, let $x_i = (\theta_{\Delta_{i1}}, s\alpha_{\Delta_{i1}}, \dots, \theta_{\Delta_{ih}}, s\alpha_{\Delta_{ih}})$ be a descriptor for each frame i , where $i = 1, \dots, M$, M is the number of frames, h the number of considered triangles, and s a scale factor that puts into balance the influence of areas over angles. Then, the descriptor X for each sequence of M frames is defined as $X = (x_1, \dots, x_M)$ and its time-stamped bag representation is the set $B_X = \{b_1, b_2, \dots, b_M\}$, where $b_i = (x_i, \beta_{\frac{i}{M}})$ is the ordered pair obtained by appending a single value to descriptor x_i . The appended

value $\beta \frac{i}{N}$ adds temporal information to the original descriptor. It is computed by multiplying the relative temporal position of the descriptor (i.e. the video frame at which the descriptors is located divided by the total number of frames in the video) by a constant β , that regulates the weight of the temporal information. We refer to β as the frame weight and to b_i as a time-stamped descriptor.

Figure 2 illustrates how the Hausdorff distance between two bags of descriptors is obtained. The compared bags correspond to the *happy* (top and middle) and *sadness* (bottom) facial expressions. Descriptors are linked to their nearest neighbor by arrows. Values at the beginning of each downward solid arrow describe the euclidean distance from one expression to the nearest expression below it, while values at the beginning of each upward dotted arrow describe the euclidean distance from one expression to the nearest above it. Observe that each descriptor is represented by its corresponding face with the selected 16 landmarks marked with dots. Moreover, faces are shown in the order they appear in the original sequence. The leftmost image of each sequence corresponds to the neutral expression and it evolves along the sequences ending in the expression itself, shown in the rightmost images in Fig. 2. It can be seen that descriptors corresponding to these neutral expressions are relatively close to each other. This is an expected behavior, since the neutral face is common to all the expressions.

On the other hand, when comparing sequences that contain different expressions, the last frames for each sequence lead to larger distances, see middle and bottom sequences in Fig. 2. This example illustrates an important assumption of our approach: even though two differently labeled expressions sequences may (and most probably will) contain many similar descriptors, at least one of the sequences will contain expressions that differentiate it from the other sequence. Therefore, when comparing the bag representations of those sequences, descriptors originated in such differentiating expressions will be at large distance from their nearest neighbors in the other bag. As a consequence, the Hausdorff Distance will be large.

Recall from Sect. 3 that the modified Hausdorff Distance can be tuned by specifying the aforementioned parameter $0 \leq K \leq 1$. To illustrate the effect of this parameter, Fig. 2 (top and middle) shows the directed Hausdorff Distance between two *happy* bags. More precisely, if we use the names A and B to denote the bags from the top and the middle respectively, the figure illustrates $h_S(A, B)$ with solid lines and $h_L(B, A)$ with dotted lines. Recall from Sect. 3 that the value of S is determined by K as $S = \max(\lfloor 5K \rfloor, 1)$. Note that the third expression in A corresponds to a *surprise* expression (considered here as noise). As a consequence, its associated descriptor lies far away from every descriptor in B (distance equal to 322.13 from its nearest descriptor). Therefore, if we set $K = 1$ (and thus $S = 5$) an undesirably large Hausdorff Distance is obtained. The desired behavior can be obtained, by letting $K = \frac{4}{5}$ (and $S = 4$). It might seem that reducing K is always beneficial. However, note that setting $K = \frac{2}{5}$ (and $S = 2$) would ignore the three more discriminative expressions in A .

4 Experiments and Results

In this Section the geometric features are tested by means of the Citation-kNN classification method. We performed several experiments for evaluating the accuracy of emotion recognition. A leave-one-out subject cross validation methodology was used to assess the performance in the study: each sequence belonging to a subject in the database is classified (i.e., tested) and the rest of the subjects are used for training.

In our experiments we utilized the extended Cohn-Kanade dataset (CK+) [3], 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. We utilized all the subjects (118) and all of its sequences (327), that are labeled with one of the 7 emotions provided by the database.

In Citation-kNN, the time-stamped versions of the features are used. They are grouped into bags and the classification process is as depicted in the previous section. Different parameter values are considered in the following ranges: $R \in [2, 20]$, $C \in [2, 20]$, $K \in [0.1, 1]$ with step 0.1, $\beta \in [1, 70]$ with step 5 and $s \in [2.9, 3.6]$ with step 0.1. Best results, shown in the last row of Table 2, were obtained for $R = 8$ for the R -nearest references, $C = 20$ for the C -nearest citers, $K = 0.9$, $\beta = 30$, and $s \in [2.9, 3.6]$. The overall performance achieved by this method is 89.30% (well-classified sequences from a total of 327).

We notice that the Citation-kNN method has the advantage that the location of the apex frame (i.e., the frame where the expression itself is achieved) can be determined automatically. As shown in Fig. 2, if we consider the distance between apex frames we observe that it increases considerably. On the other hand, the lowest distance between frames from different videos is the one that corresponds to the neutral expression.

Table 1. Accuracy with the different trained and tested datasets, OA: Onset-Apex, OAO: Onset-Apex-Offset.

Method	Train: OA Test: OA	Train: OAO Test: OAO	Train: OA Test: OAO	Train: OAO Test: OAO
Accuracy (%)	89.30	83.18	83.79	89.30

We performed an experiment to test the robustness with respect to the position of the apex in the sequence. For that, the CK+ sequences were mirrored and appended at the end of the regular sequences in order to build a longer sequence where the apex is at the center, we call this sequence OAO: Onset-Apex-Offset.

Regular CK+ sequences, called OA (Onset-Apex), were used for training and OAO sequences were used for testing, and viceversa. Also, OAO sequences were used for training and testing.

In all the cases, a leave-one-out subject cross validation methodology was employed. We observed that when the apex location in the test sequences is different from the apex location in the training sequences (for example, one is in the middle and the other at the end, respectively), the accuracy remains similar and maintains the 89.3% accuracy when the extended sequences were used for training and testing, see Table 1.

For the sake of comparison among the existing methods in the bibliography Table 2 shows the classification rates of the seven expression of CK+ dataset. All the methodologies are sequence based multi-class recognition systems. Some of the methods compared use different number of folds when performing cross-validation. In Cov3D [10] and Interval Temporal Bayesian Network (ITBN) [11] tests were performed using 5 and 15 fold-cross validation, respectively. The Constrained Local Method (CLM) [12] and Lucey’s work [3] use leave-one-out testing approach. The proposed methodology, Cit-kNN (last row in Table 2), achieves classification accuracy comparable to the state-of-the-art supervised 2D facial expression recognition techniques on CK+ database.

Table 2. Benchmark results. Best scores are in bold and 2nd best scores are underlined.

	An	Co	Di	Fe	Ha	Sa	Su
Cov3D [10]	94, 4	100, 0	<u>95, 5</u>	90, 0	96,2	70,0	100, 0
SPTS+CAPP [3]	75,0	<u>84,4</u>	94,7	65,2	100, 0	68,0	96,0
ITBN [11]	<u>91, 1</u>	78,6	94,0	<u>83, 3</u>	89,8	76, 0	91,3
CLM [12]	70,1	53,4	92,5	72,1	94,2	45,9	93,6
Cit-kNN	84,4	77,8	96, 6	44,0	<u>98, 6</u>	<u>75, 0</u>	<u>98, 8</u>

5 Conclusions

In this paper we have presented a simple geometric descriptor for facial expression recognition based on angles and areas of selected landmarks corresponding to strategic locations based on motion and position. An adaptation of the k-NN approach to the Multiple Instance Learning task has been used as a classifier. In case there is no movement in the sequence, the behavior of Citation-kNN is not affected by this fact. Also, Citation-kNN can automatically obtain the expression apex frame and performs robustly when the apex frame is located at different positions in the sequence video.

References

1. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. *IEEE Trans. PAMI* **37**, 1113–1133 (2015)

2. Zen, G., Porzi, L., Sangineto, E., Ricci, E., Sebe, N.: Learning personalized models for facial expression analysis and gesture recognition. *IEEE Trans. Multimed.* **18**, 775–788 (2016)
3. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The CK+ Dataset: a complete dataset for action unit and emotion-specified expression. In: *CVPRW*, pp. 94–101 (2010)
4. Acevedo, D., Negri, P., Buemi, M.E., Mejail, M.: Facial expression recognition based on static and dynamic approaches. In: *ICPR*, pp. 4124–4129 (2016)
5. Bottino, A., Vieira, T., Ul Islam, I.: Geometric and textural cues for automatic kinship verification. *Int. J. Pattern Recogn.* **29** (2015)
6. Osman Ali, A., et al.: Age-invariant face recognition using triangle geometric features. *Int. J. Pattern Recogn. Artif. Intell.* **29**, 1556006 (2015)
7. Corneanu, C.A., et al.: Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. *IEEE Trans. Pattern Anal.* **38**, 1548–1568 (2016)
8. Ubalde, S., Gómez-Fernández, F., Goussies, N.A., Mejail, M.: Skeleton-based action recognition using citation-kNN on bags of time-stamped pose descriptors. In: *ICIP*, pp. 3051–3055 (2016)
9. Wang, J., Zucker, J.D.: Solving the multiple-instance problem: a lazy learning approach. In: *Proceedings of the 17th International Conference on Machine Learning ICML 2000*, pp. 1119–1126 (2000)
10. Sanin, A., Sanderson, C., Harandi, M., Lovell, B.: Spatio-temporal covariance descriptors for action and gesture recognition. In: *WACV*, pp. 103–110 (2013)
11. Wang, Z., Wang, S., Ji, Q.: Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In: *CVPR*, pp. 3422–3429 (2013)
12. Chew, S.W., Lucey, P., Lucey, S., Saragih, J., Cohn, J., Sridharan, S.: Person-independent facial expression detection using constrained local models. In: *FG 2011*, pp. 915–920 (2011)