Spatio-Temporal Reasoning for Reliable Facial Expression Interpretation

J. Orozco*, F.A. García†, J.L. Arcos† and J. Gonzàlez‡

* Computer Vision Center (CVC – UAB)
Campus UAB, O-08193 Bellaterra, Spain.
† Instituto de Investigación en Inteligencia Artificial (IIIA – CSIC)
Campus UAB, E-08193 Bellaterra, Spain.
‡ Institut de Robòtica i Informàtica Industrial (UPC – CSIC),
C. Llorens i Artigas 4-6, 08028, Barcelona, Spain.

Abstract. Understanding human emotions has received contributions from the image analysis and pattern recognition areas. The most popular facial expression classifiers deal with eyebrows and lips while avoiding the eyelids. According to psychologists, eye motion is relevant for trust and deceit analysis as well as for dichotomizing near facial expressions. Unlike previous approaches, we include the eyelids by constructing an appearance-based tracker. Subsequently, a Case-Based Reasoning approach is applied by training a case-base with seven facial actions. We classify new facial expressions with respect to previous solutions by assessing the confidence for the proposed solutions. Therefore, the proposed system yields efficient classification rates comparable to the best previous facial expression classifiers. The ABT and CBR combination provides trusty solutions by evaluating the confidence of the solution quality for eyebrows, mouth and eyes. Consequently, this method is robust and accurate for facial motion coding, and for confident classifications. The training is progressive, the quality of the solution increases with respect to previous solutions and re-training processes is not required.

1 Introduction

Inner emotions are expressed through spontaneous or predetermined body and facial gestures. Therefore, many psychologists and computer science researchers deal with facial expression recognition as a communicative function of emotions.

Ekman and Friesen [7] have described facial movements with a set of action units (AUs) by developing the Facial Action Coding System. By tracking facial features and measuring facial movement, they attempt to categorize different facial expressions. Emphasis is placed on the importance of the analysis of local facial actions, as well as on the analysis of subtle facial movements.

Facial expression recognition attempts to classify the temporal deformation of faces into abstract classes that are purely based on visual information. The goal is to detect the facial actions and their intensity for a posterior classification.

Based on the work of Ekman, several approaches have been developed in order to recognize facial expressions by using classification techniques, such as Neural

Networks, Gabor Wavelets, Bayesian Networks, LDA, Support Vector Machines, Neighbour Networks, etc [8]. All of them differ in robustness, accuracy, number of training examples and effectiveness. They report an average effectiveness of 86%, depending on training, high expressiveness of the emotions, and new subjects.

We propose a new approach by combining Appearance-Based Trackers (ABT) and Case-Based Reasoning (CBR). ABT allows encoding facial actions by providing the temporal deformation of the face [6]. Firstly, we propose an ABT providing seven facial actions, which correspond to eyebrows, lips and eyelids. The novelty of our approach is that we are able to include eyelid information, while previously it was not possible due to the difficulty of the blink motion. Secondly, a Case-Based Reasoning system (CBR) is trained with seven facial expressions from standard databases. This technique classifies new expressions by using previously solved cases, which are stored in a case-base and categorized by problem descriptions and solutions [1,5].

CBR has a low cost of knowledge acquisition by recording new useful cases. The system retrieves the most similar face configuration in order to apply the past solution to the new problem. Consequently, knowledge maintenance is straight-forward because the system learns incrementally and the solution quality is increased even when the domain is ill defined. The problem-solving efficiency is enhanced by evaluating the confidence of the solution [9].

In that way, the proposed approach allows a spatio-temporal classification for reliable facial expressions. Our system is trained with seven facial expressions, which are encoded by applying appearance-based trackers. Next, we apply a k-NN classifier in order to identify the possible solutions, which are posteriorly evaluated with confidence predictors. CBR is suitable for evolutionary learning, which is beyond the scope of this paper.

The paper is organized as follows: section 2 describes the facial feature extraction method and the appearance-based tracker. Section 3 explains the case-based reasoning approach for recognizing facial expressions. Section 4 presents experimental results and discussion. Finally, section 5 concludes the paper.

2 Appearance-Based Tracker

In order to extract facial movements from eyebrows, lips and eyelids, we use the 3D face Candide Model, which provides a simple process to construct an appearance model and a single parametrization to extract facial features [6]. The 3D face model is given by the three spatial coordinates of each vertex. The shape is described by the $(n \text{ x } i)$ matrix $\mathbf{F}$, where n is the number of vertices and i is their coordinates:

$$\mathbf{F}_n^i = \mathbf{f}_n^i + \mathbf{D}_n^{i,d}\vartheta_d + \mathbf{E}_n^{i,e}\gamma_e, \tag{1}$$

where $\mathbf{f}$ is the standard configuration, $\mathbf{D}$ encodes the biometry of each person, and $\mathbf{E}$ handles the facial animations. We consider $d = 16$ biometric parameters encoded by the control vector ϑ, as well as $e = 7$, the facial action parameters

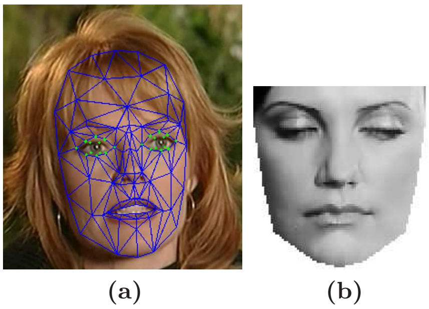


Fig. 1. **(a)** The 3D mesh is projected onto the input image I_t in order to construct the appearance model $\mathbf{A}_t$ by a warping process **(b)**.

encoded by the control vector γ, which corresponds to eyebrows, lips and eyelids, see Fig. 1.(a).

We extract the tracking vector as $\mathbf{q}$, which contains the head pose (three Euler's angles, scale and image coordinates) and the animation parameters. The tracking vector is as follows:

$$\mathbf{q} = [\theta_x, \theta_y, \theta_z, s, t_x, t_y, \gamma_0, ..., \gamma_6] \;=\; [\boldsymbol{\alpha}, \boldsymbol{\gamma}]. \tag{2}$$

Given an image sequence $\mathbf{I}_t$, depicting head motion and facial expressions, we model each face by constructing an appearance-based model, which projects the 3D mesh onto the input image for a specific configuration of the vector $\mathbf{q}$, see Fig. 1.(b). Therefore, for each input image, the goal is to estimate the vector $\mathbf{q}$, which gives us seven facial actions for eyebrows, lips and eyelids.

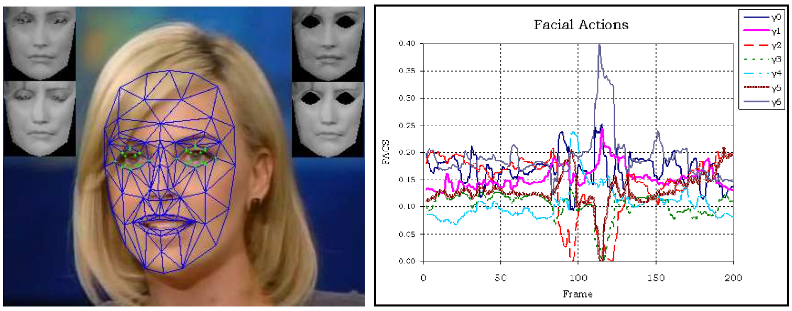
2.1 Facial Action Extraction

In order to estimate the corresponding vector $\mathbf{q}_t$ at each frame, we construct the corresponding appearance model, $\mathbf{A}_t$, by applying a warping process, $\Psi(\mathbf{I}_t, \mathbf{q}_t) \rightarrow \mathbf{A}_t(\mathbf{q}_t)$. Consequently, the appearance model depends on the vector $\mathbf{q}_t$ and the animation parameters $\boldsymbol{\gamma}_t$.

Each appearance model $\mathbf{A}_t$ is assumed to follow a Gaussian distribution, $N(\mu_t, \sigma_t)$. Therefore, we can apply a filtering technique that is time efficient for estimating the Gaussian parameters over time with respect to previous estimations. All estimated appearances, $\hat{\mathbf{A}}$, are held under an exponential with an updating factor ω as follows;

$$\begin{aligned} \mu_{t+1} &= \omega\mu_t + (1-\omega)\hat{\mathbf{A}}_t, \\ \sigma^2_{t+1} &= \omega\sigma^2_t + (1-\omega)(\hat{\mathbf{A}}_t - \mu_t)^2 \end{aligned} \tag{3}$$

An adaptive velocity model is adopted in order to estimate the vector $\mathbf{q}_t$. The current input image $\mathbf{I}_t$ is registered with the current appearance model,



(a) **(b)**

Fig. 2. Multi-tracking system **(a)** providing seven facial actions (FACS) for an input image sequence **(b)**.

$\mathbf{A}_t$, which depends on the estimated vector $\hat{\mathbf{q}}$. The final estimation is obtained by minimizing the Mahalanobis distance between the estimated and the current average appearances. Here, the appearance parameters μ and σ are known, and the distance is minimized by an iterative first-order linear approximation and calculating the Jacobian matrix:

$$\mathbf{A}_t \approx \hat{\mathbf{A}}_{t-1} + \frac{\partial(\mathbf{A}_t, \mathbf{q}_t)}{\mathbf{q}_t}(\mathbf{q}_t - \hat{\mathbf{q}}_{t-1}), \; and$$

$$\Delta\mathbf{q}_t = \mathbf{q}_t - \hat{\mathbf{q}}_{t-1} = -\mathbf{J}_t^*[\Psi(\mathbf{I}_t, \hat{\mathbf{q}}_{t-1}) - \mu_t]. \tag{4}$$

where $\mathbf{J}_t^*$ is the pseudo inverse of the Jacobian matrix. Thus, we apply a gradient descent method by partial differences, which is able to accommodate appearance changes while achieving precise estimations.

Since eyelid motion is a faster facial action than eyebrows and lips, and the eye region motion adds uncertainty to the appearance model, another tracker is constructed in order to correct the head pose estimations. Subsequently, two ABTs are combined hierarchically; the first tracker **A**, is devoted to tracking the head, eyebrows, lips and eyelids. The second one **A'**, enhances the estimations of the first tracker by correcting the head pose, see Fig. 2.

We encode seven facial actions, which are the data for the expression classifier. Eye motion analysis provides relevant information that is useful to recognize similar facial expressions. This information has not been considered previously due to the difficulty of capturing the eye blinking. However, our system achieves robust and accurate estimations that are suitable for real-time systems, an average of 21 frames per second for each image.

We first estimate the vector **q** by obtaining the appearance model $\mathbf{A}_t$. This ABT estimates eyebrows, lips and eyelids. Subsequently, the second tracker **A'**

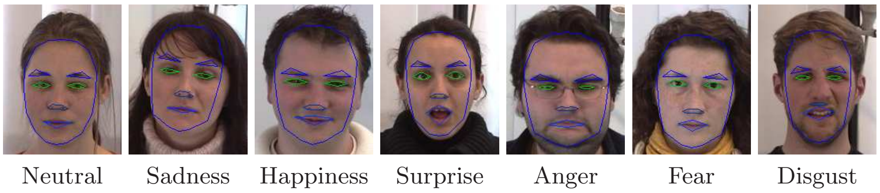


Fig. 3. Facial Expressions from FGnet, http://www-prima.inrialpes.fr/FGnet/

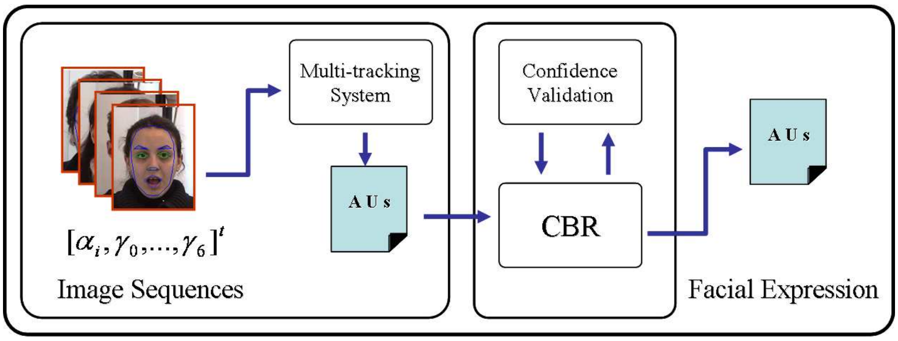


Fig. 4. Facial Expression Recognition System.

corrects the previous estimations by avoiding the eye region information and applying a small gradient descent and providing the final expressions, see Fig. 3.

3 CBR Problem-Solving

Case-based reasoning is a learning approach that imitates the way humans solve problems by means of reasoning about current situations and re-using previous solved problems. The problems, in the CBR approach, are called cases and correspond to the training data for the case-base database as well as the testing data, see Fig. 4. The new problem is solved by retrieving cases with similar attributes to the target case in order to classify by re-using the previous solutions.

Consequently, the development of CBR systems has increased the necessity to support the analysis of the case-base structure while providing solutions with a required accuracy. Some CBR approaches handle the solution quality by using as similarity criteria the amount of cases that indicate confidence [11,9,4]. According to Cheetham and Price, CBR systems should be able to attach a solution confidence while estimating the problem's solutions and evaluating accuracy for these estimations [3,2]. As a result, the CBR system can estimate both a problem solution and an accuracy evaluation for the proposed solution.

(case:**id** fgnet happy2	
:**Problem-Description**	
:upper lip riser	0.169
:jaw drop	0.137
:lip stretcher	0.123
:brow lowerer	0.122
:lip corner depressor	0.121
:outer brow raiser	0.121
:eyelid riser	0.207)
:**Solution-Description**	'happy)

Table 1. Case-base description.

In order to recognize facial expressions, a CBR system follows the CBR cycle, which consists of the following four steps: *Retrieve* the most similar cases to the current problem. *Reuse* the information and the knowledge of the retrieved cases for proposing solutions. *Revise* the solution by assessing the confidence of the proposed solution. *Retain* the new solved problem for posterior problem solutions [1]. Each case or problem is composed of *identification*, *problem-description* with seven normalized facial actions, and *solution-description*, see Tab. 1. The classification process considers only the facial actions while avoiding the head pose. Our contribution enhances the first two steps of the CBR cycle as follows:

Case-Retrieve: The goal of the retrieve step is to find the set of possible solutions for a given input image $\mathbf{I}_t$, which is encoded by the multi-tracking system. This system provides the seven facial actions (FACS) that constitute the problem description. Therefore, the problems can be compared by considering the Euclidean distance, D, between the input facial expression $\gamma = [\gamma_0, ..., \gamma_6]$ and the *problem-description* for each case from the case-base. Consequently, the k-most similar cases RC to the target problem are chosen by applying k-NN.

Case-Reuse: We evaluate confidence for the target case with respect to the previous retrieved cases RC in order to set a confidence threshold relative to k.

These measures are applied as confidence predictors by assuming that both the amount of neighbours and the similarity relationship among cases, play an important role in the assessment of the predictors. We explain the confidence predictors in the next section.

3.1 Confidence Predictors

We assess the confidence of the solution by computing five predictors based on k-NN classification at the *Case Revise* step [5]. We consider the following confidence predictors:

1. *AvgNUNIndex*: The Average Nearest Unlike Neighbour Index (different class to the target case) measures how close are the first k NUN_s to the target case, c:

$$AvgNUNIndex(c,k) = \sum_{i=1}^{k} IndexOfNUN_i(c). \tag{5}$$

2. *SimRatio*: The Similarity Ratio calculates the relative ratio of the similarity between the target case c and its k Nearest Like Neighbours (the same class to the target case), NLN_s with respect to the similarity between the target case c and its k NUN_s:

$$SimRatio(c,k) = \frac{\sum_{i=1}^{k} Sim(c, NLN_i(t))}{\sum_{i=1}^{k} Sim(c, NUN_i(c))}. \tag{6}$$

3. *SimRatio'*: The Similarity Ratio within k is similar to the above measure except that it only uses the NLN_s and NUN_s from the first k neighbours:

$$SimRatio'(c,k) = \frac{\sum_{i=1}^{k} Sim(c, NN_i(c))1(c, NN_i(c))}{1 + \sum_{i=1}^{k} Sim(c, NN_i(t))(1 - 1(c, NN_i(c)))}. \tag{7}$$

4. *Sum of NN Similarity* is the total similarity of the NLN_s in the first k neighbours of the target case c:

$$SumNNSim(c,k) = \sum_{i=1}^{k} 1(c, NN_i(c))Sim(c, NN_i(c)). \tag{8}$$

5. *AvgNN Similarity* is the average similarity of the NLN_s in the first k neighbours of the target case c:

$$SumNNSim(c,k) = \frac{\sum_{i=1}^{k} 1(c, NN_i(c))Sim(c, NN_i(c))}{\sum_{i=1}^{k} 1(c, NN_i(c))}, \tag{9}$$

where $NN_i(c)$ denotes the i^{th} nearest neighbour of the case c. As well, $NLN_i(c)$ is the i^{th} nearest like neighbour and $NUN_i(c)$ is the i^{th} nearest unlike neighbour of the case c. $1(a,b)$ denotes if the case a belong to the same class of the case b. The effectiveness of each measure depends on the proportion of cases correctly predicted. The highest confidence is obtained when the number of incorrect predictions is zero. The goal of this calculation is to obtain a confidence value. We evaluate each measure with different values of k in order to set a threshold value for each predictor. Subsequently, the final confidence value is obtained by means of a weighted-sum with the above threshold values, see Fig. 5.

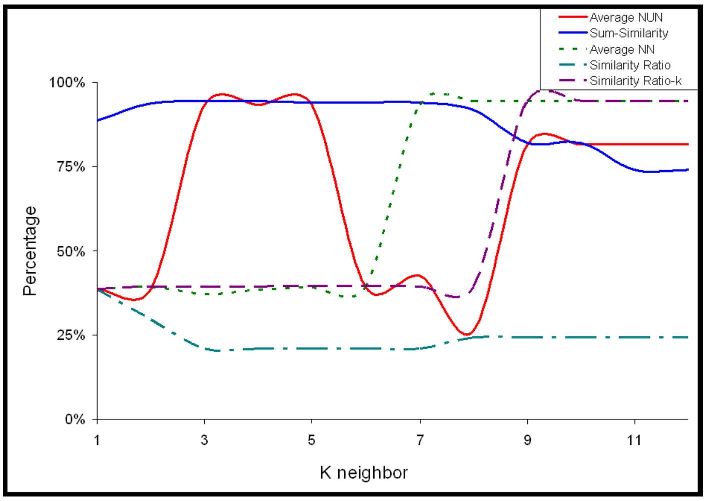


Fig. 5. The five predictors are compared in order to set a confidence threshold.

4 Experimental Results

We have used the FGnet database, which contains facial image sequences of the basic seven emotions defined by Ekman and Friesen [7]. We have chosen different kind of spontaneous faces by taking into account gender, skin colour and expressiveness. The case-base (CB) was trained with 543 cases (facial expressions) corresponding to three actors for each expression, where an actor could be included in more than one expression. The label for each expression is provided by the database and the number of images per expression depends on the time required to get the peak of the expression. As well as the 543 facial expressions used for *training*-CB, 980 additional were used for testing, *testing*-CB.

The input data for the CBR system is provided by the multi-tracking system. These facial expressions are standardized into the range [0,1] in order to apply probabilistic metrics. The provided labels by FGnet are the corresponding solutions for the facial expressions for the *training*-CB process, see Table 1.

In the *training*-CB, we apply a leave-one-out process [10] to set the confidence threshold. Firstly, given a target case extracted from case-base CB, k-NN is applied and the similarity measure used is Euclidean distance. Secondly, the confidence is estimated for the target case by means of agreement of the five predictors for each of the retrieved-cases RC. Finally, the above process is done iteratively for $k = 1, ..., 12$ for each case in order to set the relationship between the confidence assessment and the k value, see Fig. 5. Subsequently, we decide $k = 9$ as the optimum value for obtaining the highest values of the confidences.

In the *testing*-CB process, we compared the classification rate for 980 expressions for three different experiments. Firstly, we applied k-NN without including the eye motion while obtaining an average effectiveness of 86%. Secondly, we

Emotion	Neutral	Happy	Anger	Disgust	Fear	Sad	Surprise	Effectiveness	# Images
Neutral	93	0	3	0	3	0	0	93.33%	30
Happy	2	95	1	1	0	2	0	95.49%	133
Anger	2	1	93	0	1	1	1	93.33%	135
Disgust	2	1	0	94	1	2	1	94.31%	123
Fear	1	0	1	0	96	1	0	96.32%	136
Sad	0	2	2	2	3	89	2	89.00%	300
Surprise	1	0	1	0	1	0	98	97.56%	123
								94.19%	980

Table 2. Confusion Matrix by assessing Confidence in the $Testing-CB$ process.

have included the eye motion by applying k-NN and the optimum k confidence threshold but without assessing confidence. This improves the classification effectiveness to 90%. Finally, we evaluated the confidence for the proposed solution, which improves further the classification to 94%. These results prove how the quality of the solution improves by using the optimum k-confidence-threshold and assesing the confidence for the solved problem, see Table. 2.

As a result, the classification effectiveness increases by improving the solution quality with the confidence assessment, which reclassifies the misclassified cases. The misclassification rate depends on the nearness of clusters. The FGnet database has image sequences of spontaneous expressions, where the actor starts with a neutral expression until the peak of the expression. Subsequently, we have at the beginning of each sequence, similar configurations of the vector $\gamma = [\gamma_0, ..., \gamma_6]$, which results in the nearness among classes.

Ekman also stated that neutral and sadness expressions are highly similar if the subtle eyelid motion is not considered. Therefore, our system was tested without including eye motion as a problem attribute and the confidence assessment we obtained an average classification rate of 86% comparable to previous approaches in the literature. But a classification rate of 94.2% was achieved by assessing the confidence and including the eye motion in the testing process.

5 Conclusions

Due to the necessity for robust and accurate methods for facial expression evaluation, as well as for dynamic classification methods, we proposed a facial expression recognition system that applies case-based reasoning and appearance-based trackers. The facial movements are extracted by using ABTs for providing the problem descriptions for the case-base. Therefore, the eyelid motion information increases the classification effectiveness while improving previous classifiers.

CBR constitutes a spatio-temporal reasoning system able to classify facial expressions while improving previous classifiers. We set the optimum k-confidence-threshold by applying a leave-one-out process. Next, we improve the effectiveness by assessing confidence proposed solutions and achieving a 94% of effectiveness.

The confidence evaluation and eyelid motion analysis allow improving the separability of facial expression clusters.

Future work will include gaze analysis and knowledge maintenance for emotion evaluation. Also, learning and aggregation methods for CBR will be considered and context information for a posterior cognitive emotion evaluation.

Acknowledgement:
This work is supported by EC grants IST-027110 for the HERMES project, IST-045547 for the VIDI-Video project, by the Spanish MEC under projects TIN2006-14606 and DPI-2004-5414. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, 7(1):39–52, 1994. (Cited on pages 2 and 6)
2. W. Cheetam and J. Price. Measure of a solution accuracy in case-based reasoning systems. *In Funk, P.,González-Calero, P.,eds.: 7th European Conference on Case-Based Reasoning (ECCBR 2004)*, 3155:106–118, 2004. (Cited on page 5)
3. W. Cheetham. *Case-Based Reasoning with Confidence*. Book Advances in Case-Based Reasoning, 2000. (Cited on page 5)
4. J. Chua and P. Tischer. Determining the trustworthiness of a case-based reasoning solution. *In International Conference on Computational Intelligence for Modeling, Control and Automation:7th European Conference on Case-Based Reasoning (ECCBR 2004)*, 3155(1740881885):12–14, 2004. (Cited on page 5)
5. S. Delany, P. Cunningham, and D. Doyle. Generating of classification confidence for a case-based spam filter. In *In 6th International Conference on Case-Based Reasoning (ICCBR 2005).Springer*, volume 3620, pages 177–190, 2005. (Cited on pages 2 and 6)
6. F. Dornaika, J. Orozco, and J. Gonzàlez. Combined head, lips, eyebrows, and eyelids tracking using adaptive appearance models. In *In 4th International Conference on Articulated Motion and Deformable Objects (AMDO 2006).Springer*, volume 4069, pages 110–119, 2006. (Cited on page 2)
7. P. Ekman and W. V. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press, Palo Alto*, 1978. (Cited on pages 1 and 8)
8. B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(2003):259–275, 2003. (Cited on page 2)
9. M. Grachten, F.A. García, and J.L. Arcos. Navigating through case base competence. In *7th European Conference on Case-Based Reasoning (ECCBR 2004), Springer*, volume 3155, pages 282–295, 2004. (Cited on pages 2 and 5)
10. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Conference 18 on Artificial Intelligence (IJCAI)*, pages 1137–1145, 1995. (Cited on page 8)
11. T. Reinartz, I. Iglezakis, and T. Roth-Berghofer. Computational intelligence. *Artificial Intelligence Communications*, 2001(17):214–234, 2001. (Cited on page 5)