Synthesis of Facial Expressions in Photographs: Characteristics, Approaches, and Challenges

RAFAEL LUIZ TESTA, CLÉBER GIMENEZ CORRÊA, ARIANE MACHADO-LIMA, and FÁTIMA L. S. NUNES, University of São Paulo, Brazil

The synthesis of facial expressions has applications in areas such as interactive games, biometrics systems, and training of people with disorders, among others. Although this is an area relatively well explored in the literature, there are no recent studies proposing to systematize an overview of research in the area. This systematic review analyzes the approaches to the synthesis of facial expressions in photographs, as well as important aspects of the synthesis process, such as preprocessing techniques, databases, and evaluation metrics. Forty-eight studies from three different scientific databases were analyzed. From these studies, we established an overview of the process, including all the stages used to synthesize expressions in facial images. We also analyze important aspects involved in these stages such as methods and techniques of each stage, databases, and evaluation metrics. We observed that machine learning approaches are the most widely used to synthesize expressions. Landmark identification, deformation, mapping, fusion, and training are common tasks considered in the approaches. We also found that few studies used metrics to evaluate the results, and most studies used public databases. Although the studies analyzed generated consistent and realistic results while preserving the identity of the subject, there are still research themes to be exploited.

CCS Concepts: • **Computing methodologies** → **Image-based rendering**;

Additional Key Words and Phrases: Facial expression synthesis, facial expression mapping, facial expression cloning, expression transfer, facial expression generation

ACM Reference format:
Rafael Luiz Testa, Cléber Gimenez Corrêa, Ariane Machado-Lima, and Fátima L. S. Nunes. 2019. Synthesis of Facial Expressions in Photographs: Characteristics, Approaches, and Challenges. *ACM Comput. Surv.* 51, 6, Article 124 (January 2019), 35 pages.
https://doi.org/10.1145/3292652

1 INTRODUCTION

The recognition of facial expressions by human beings is a skill that contributes to social interaction (Hess 2001). However, the presence of some disorders causes individuals to present deficiencies in this ability, hindering communication and other activities of the human daily routine (Rocca et al. 2009). Examples of these disorders are autistic spectrum disorders (Harms et al. 2010), mood disorders (Rocca et al. 2009), and schizophrenia (Taylor and MacDonald 2012). Research studies indicate that people with some of these disorders can be trained to improve their ability to recognize emotions (Cheng and Ling 2008; Golan and Baron-Cohen 2006; Grynszpan et al. 2008;

Authors' address: R. L. Testa, C. G. Corrêa, A. Machado-Lima, and F. L. S. Nunes, University of São Paulo, School of Arts, Sciences and Humanities, Rua Arlindo Bettio, 1000, São Paulo, SP, 03828-000, Brazil.

0360-0300/2019/01-ART124 $15.00
https://doi.org/10.1145/3292652

Lahiri et al. 2013), and, especially, this training can be more easily performed with faces familiar to the individual. For example, people with an autism spectrum disorder have greater difficulty in identifying emotions in unfamiliar faces (Pierce et al. 2004).

Computational tools can be used for diagnosing and treating these disorders (Grynszpan et al. 2008; Lahiri et al. 2013). For example, synthetic images can be part of a tool for training the ability to recognize emotions in facial expressions, as in some games (Cheng and Ling 2008). However, the preparation of a face image with a variety of expressions that is, at the same time, familiar to the player, is a challenging task involving several computational techniques. In general, the pixels of the human face image must be manipulated to produce the desired expression.

In addition to the training context, the synthesis of facial expressions has applications in automation of interactive web agents for video-conferencing with low bandwidth (Li et al. 2007), interactive interfaces (Mendi and Bayrak 2011), improvement of photographs (Fujishiro et al. 2009), facial surgery planning (Keeve et al. 1998), computer animation (Noh and Neumann 2001), and custom icons (Li et al. 2007).

Although the scope of this article is facial expression synthesis, the results may be useful for several related areas. For instance, facial expression recognition usually involves facial landmarks identification, features extraction, and the use of facial expression databases, topics addressed in Sections 4.1, 6.1, and 9, respectively. Facial expression synthesis techniques can also be used in less obvious ways, such as in training data augmentation (Mohammadian et al. 2016) and in inspiring new features (Song et al. 2010) for facial expression recognition. It can even be part of a face recognition system that is facial expression invariant (Amberg et al. 2008) or an age progression/regression system that considers facial expression (Zhang et al. 2017).

The universality of facial expressions of emotions is an old discussion (Aristotle et al. 1913; Darwin 1916). Since the discovery of six universal facial expressions of emotions (happiness, sadness, anger, disgust, fear, and surprise) (Ekman et al. 1972; Izard 1971), these have been commonly used in computational fields, such as facial expression recognition (Bettadapura 2012; Li and Deng 2018; Mehta et al. 2018) and synthesis (Abboud and Davoine 2004, 2005; Leung et al. 1996; Li et al. 2014; Tay et al. 2009; Wang and Wang 2008; Wei et al. 2016; Zhou and Lin 2005). Although these researchers consider these emotions as universal, this is a controversial issue (Russell 1994). In Jack et al. (2016), only a subset of the emotions (happiness, surprise, anger, and sadness) is considered universal. Another study even reports that facial displays have specific functions in social interactions and depend upon context (Crivelli and Fridlund 2018); that is, people can express and interpret emotions in different ways depending on context and culture.

In this article, we consider the set of six emotions because they are most frequently used in facial expression synthesis. Each facial expression of emotion corresponds to a specific pattern of movements of facial muscles (Ekman and Friesen 1971). The minimum movements of a muscle or group of muscles involved in a facial expression of emotions were defined by Ekman and Friesen (1976) as *action units*. It is possible to reproduce these patterns of movement and thus simulate each facial expression by manipulating the image's content.

The simulation of facial expressions in computerized systems has been addressed in many ways, such as through generating caricatures (Testa et al. 2015; Yang et al. 2009), three-dimensional (3D) reconstructions (Blanz et al. 2003; Liang et al. 2016; Shu et al. 2013), avatars (Boker et al. 2009; Saragih et al. 2011a; Zhang et al. 2010), photographs (Li et al. 2014; Xie et al. 2015; Zhang et al. 2003, 2014), and movements of the muscles of a robotic head (Moosaei et al. 2015; Wu et al. 2009). There is a fair number of studies in the literature, spread over several databases, which makes it difficult to compose an overview of the area. Therefore, this study aims to compile those studies, limiting its scope to studies whose objective is the synthesis of facial expressions in photographs.

Ersotelos and Dong (2008) presented a literature review in which they report the construction of realistic face images from two perspectives: 3D modeling and animation/dynamic synthesis. For each of the two perspectives, the review categorizes the approaches according to similarities in their methods. The objective of the review is similar to the present study. However, our Systematic Review (SR) addresses the subject more specifically, including the methods and techniques used to synthesize facial expressions in photographs. For these characteristics, our study aims to analyze the approaches related to Image-Based Rendering (IRB). The IRB approach synthesizes the facial expression from facial movements extracted from recorded videos or static images (Ersotelos and Dong 2008). Even though it is more specific, this review also addresses the subject in greater depth by considering the image databases of facial expressions; deformation, mapping, and machine learning approaches; the forms of evaluation of synthesized images; and the metrics used in these evaluations. In addition, this study has a larger scope of time, covering articles published until June 2018 and therefore including studies that were not available in Ersotelos and Dong (2008).

Thus, this SR aims to identify the methods and techniques existing in the literature to synthesize and evaluate facial expressions in photographs. The review was based on the following research questions:

(1) What methods and techniques are used for synthesizing facial expressions in photographs?
(2) What types of evaluation are used to check the realism of these expressions?

Section 2 of this article presents the basic concepts of SR and the elements comprising its protocol, Section 3 globally analyzes the articles included according to our SR protocol, and Section 4 discusses the preprocessing procedures used in the studies included. Later sections discuss the main approaches of summaries referred to in the literature: Section 5 discusses deformation and mapping approaches, Section 6 examines machine learning approaches, and Section 7 presents other approaches. Section 8 discusses the different methods and metrics used by the authors to evaluate synthesized images. Section 9 presents an overview of the facial expression image databases mentioned in the studies analyzed. Finally, Section 10 discusses the trends, challenges, and opportunities from the interpretation of studies included, and Section 11 presents the findings on the subject.

2 RESEARCH METHOD

This SR was followed three classical stages: (i) planning and selection of studies, (ii) data extraction, and (iii) data interpretation (Petticrew and Roberts 2008).

The researched academic databases were defined in accordance with the articles obtained from an exploratory analysis previously conducted. The following keywords were considered: "facial expression," "synthesis," and "map." The keywords "3D" and "robotic" were used to exclude studies during the search. First, searches were carried out with strings formed by these keywords in the data bases *Science Direct*[1], *IEEE Xplore*[2], and *ACM Digital Library*[3]. The search engines were requested to search the keywords in the heading, summary, and keywords of the studies, as presented in Table 1.

Table 2 presents the inclusion (I) and exclusion (E) criteria for selection of the studies of interest. An article was included if it had at least one of the inclusion criteria and excluded if it presented at least one of the exclusion criteria.

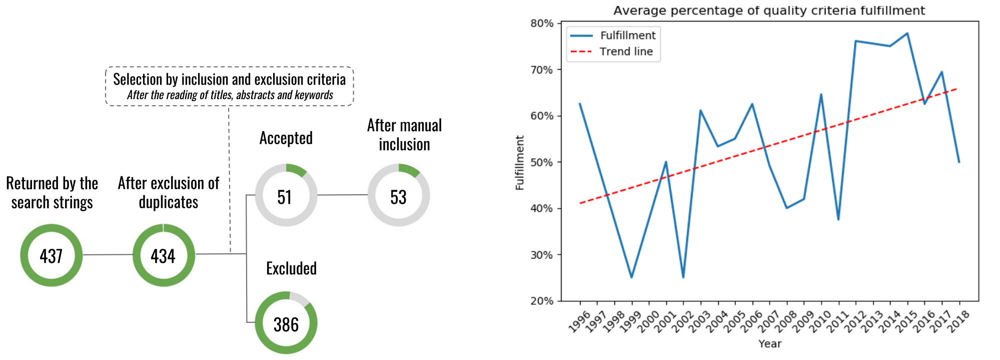
[1] http://www.sciencedirect.com/.
[2] http://ieeexplore.ieee.org/.
[3] http://dl.acm.org/.

Table 1. Strings of Search by Academic Database

Database	String
Science Direct	*TITLE-ABSTR-KEY((“facial expression” AND (synthesis OR map)) AND NOT (3D OR robot))*
IEEE Xplore	*(“Document Title”:“facial expression” OR “Abstract”:“facial expression” OR “Author Keywords”:“facial expression”) AND (“Document Title”:synthesis OR “Abstract”:synthesis OR “Author Keywords”:synthesis OR “Document Title”:map OR “Abstract”:map OR “Author Keywords”:map) NOT (“Document Title”:3D OR “Abstract”:3D OR “Author Keywords”:3D OR “Document Title”:robot OR “Abstract”:robot OR “Author Keywords”:robot)*
ACM Digital Library	*(acmdlTitle:(“facial expression”) OR recordAbstract:(“facial expression”) OR keywords.author.keyword:(“facial expression”)) AND ((acmdlTitle:(synthesis) OR recordAbstract:(synthesis) OR keywords.author.keyword:(synthesis)) OR (acmdlTitle:(map) OR recordAbstract:(map) OR keywords.author.keyword:(map))) AND (acmdlTitle:(-3D -robot) AND recordAbstract:(-3D -robot) AND keywords.author.keyword:(-3D -robot))*

Table 2. Inclusion and Exclusion Criteria

Criterion	Description
I1	Studies that address the complete or partial synthesis of facial expressions in images will be included.
I2	Studies that describe evaluation methods regarding the synthesis of facial expressions will be included.
E1	Studies in which the synthesis occurs in images that are not two-dimensional will be disregarded.
E2	Studies in which the synthesis does not occur in photographs will be disregarded.
E3	Studies in which the synthesis occurs in videos or use videos in their methodology will be disregarded.
E4	Studies that address only speech synthesis will be disregarded.
E5	Studies that use motion sensors in the synthesis of the images will be disregarded.
E6	Duplicate studies will be disregarded.



(a) Summary of steps taken during SR.

(b) Graph of the average frequency of the fulfillment of the quality criteria.

Fig. 1. SR steps and quality criteria.

Figure 1(a) summarizes the steps executed during the SR. Only three articles were duplicates of the 437 articles recovered. The inclusion and exclusion criteria were applied by reading the heading, abstract, and keywords. After this step, 51 articles were selected for full reading, data extraction, and composition of the final analysis of this SR. Additionally, we manually included two publications, due to their importance and relevance to this SR, that had not been retrieved during the searching phase: Choi et al. (2017), which had not been initially found because it was not indexed by the databases used in this review, and Xie et al. (2018), which had not been retrieved

Table 3. Quality Criteria by Factor and Percentage of Fulfillment

Factor	**Criterion**	**Description**	**Fulfillment**
Source images	Q1	The study specifies the image databases used or the characteristics of the images acquired.	72% of 53
Facial landmarks	Q2	The study specifies how facial landmarks were identified.	74% of 47
	Q3	The study specifies which facial landmarks are used.	74% of 47
Synthesized images	Q4	Images synthesized by the study present the internal region of the mouth (teeth, lips, and tongue).	78% of 51
	Q5	The images synthesized by the study present wrinkles.	57% of 53
	Q6	The study shows at least one example of the synthesized images.	96% of 53
Objective evaluation of results	Q7	The study presents an objective evaluation of the synthesis of facial expressions.	26% of 53
	Q8	The objective evaluation of results compares the synthesized images with reference images (ground truth or gold standard).	25% of 16
	Q9	The objective evaluation of results compares the synthesized images with images synthesized by other studies.	44% of 16
Subjective evaluation of results	Q10	The study presents a subjective evaluation of the synthesis of facial expressions.	19% of 53
	Q11	The subjective evaluation of results presents the user's characteristics.	31% of 13
	Q12	The subjective evaluation of results shows some consistency mechanism.	31% of 13

because it is in an academic database (*Springer Link*[4]) that was not included in our searching protocol. All articles included cover the synthesis of complete or partial facial expressions, and their goal is the synthesis of facial expressions or the use of the synthesis of expressions for another task.

Next, the articles included were classified according to the quality criteria presented in Table 3. To each criterion met, the article added a point to its total score. The final classification in accordance with the score of these criteria allows us to analyze if the included articles report the elements analyzed in this review. The documents generated by the conduction of this SR are available online[5].

3 GLOBAL ANALYSIS OF ARTICLES INCLUDED

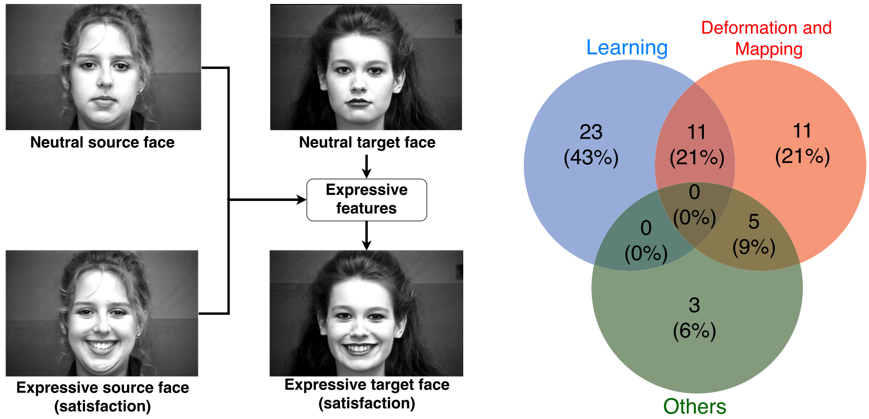
The quality criteria (Table 3) refer to the different aspects analyzed, such as specification of the source images (Section 9), detailing of facial landmarks and techniques used to extract them (Section 4.1), description of aspects related to synthesized images (Section 8.3), and detailing of factors related to objective (Section 8.1) and subjective (Section 8.2) evaluation metrics.

Table 3 also presents the percentage of articles that fulfill each quality criterion defined. This percentage corresponds to the number of articles that meet the criterion in relation to the total number of studies in which the criteria are applicable. Some quality criteria are not applicable to some articles due to their scope. For example, it is not possible to apply criterion Q4 (Table 3), which covers the internal region of the mouth, to an article whose scope is restricted to the synthesis of facial expressions for the region of the eyes, as in Xiong et al. (2010, 2007b).

The criteria related to the synthesized images presented a higher percentage of compliance probably because it referred to the main scope of the study. Conversely, the evaluation criteria

[4] https://link.springer.com.

[5] Available at: http://lapis.each.usp.br/en/sr-fes/.



(a) Definition of image types involved in the synthesis. Images taken from Lucey et al. [2010] © Jeffrey Cohn.

(b) Venn Diagram of distribution of the approaches.

Fig. 2. Image types and approaches.

presented a lower compliance rate in relation to the others, evidencing that not all the studies present systematic evaluation of results. The groups of quality criteria defined have been used to systematize the discussion of this article, as presented in Section 10. Figure 1(b) shows that the quality of the studies increased in accordance with our criteria. This figure shows the average of all criteria by year, using the same computation applied to Table 3.

Most of the studies analyzed used an image-based approach for synthesis. In this approach, the features of the facial expression made by a person (source face) are transferred to the image of the face whose expression will be reenacted (target face). The relevant features of a facial expression are identified and extracted from the source face(s). After that, these features are adapted to the target face. Finally, the extracted and adapted features are merged with the target face to synthesize the new image. The images involved in this process are exemplified in Figure 2(a).

In general, the process of facial expression synthesis is divided into preprocessing, synthesis, and postprocessing stages, as schematized in Figure 3. We only consider a category if there is more than one article that uses that approach. The preprocessing is composed of one or more of the following stages: landmark identification (Section 4.1), face alignment (Section 4.2), and source image selection (Section 4.3).

After the preprocessing stage, the synthesis stage can be performed using different approaches such as deformation and mapping (Section 5), machine learning (Section 6), and/or other methods (Section 7). Hybrid approaches are also presented in Sections 5 to 7. The classification of the approaches into machine learning and deformation and mapping was inspired and adapted from Xie et al. (2015), but we also considered the classification of approaches presented in the review Ersotelos and Dong (2008) for "other" approaches. Finally, we added new categories regarding hybrid approaches.

The studies presented different techniques seeking the best result for the synthetic images. For example, hybrid approaches construct their methods using stages from more than one approach, as can be seen in Figures 3 and 2(b). Figure 2(b) shows that 30% (16) of the studies present hybrid approaches. The overview of the stages used by each approach is presented in Figure 3, in which only the approaches that had at least two articles in the category were considered. Some studies did not have all the stages of the category, and those stages were therefore considered optional.

Some of these studies also presented a postprocessing stage to merge the synthetic expressive facial features into the source images. The techniques used at this stage included wrinkle mapping

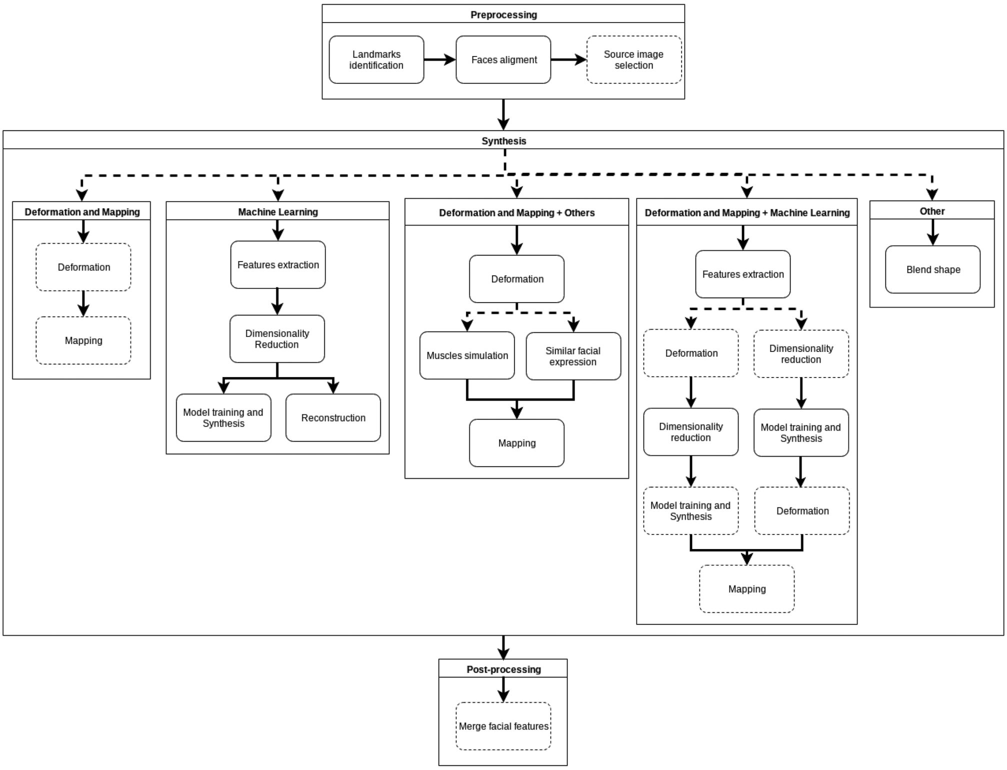


Fig. 3. Stages of the synthesis of facial expressions by approach. Dashed lines indicate an optional stage. Rectangles with dashed lines indicate optional stages, and arrows with dashed lines indicate that only one of the paths must be followed.

(improved seamless cloning) to add wrinkles to the face (Xie et al. 2017), blend ratio to join the synthesized expression with the original image (Lee et al. 2007, 2009), and fade-in-fade-out blending to prevent discontinuous regions (Xiong et al. 2009). The purpose of these processes is to improve the unification of the synthesized facial expression with the original image, aiming to provide greater realism to the result.

All the studies discuss the images synthetically generated, which will be detailed in Section 8.3. Some of them, however, go beyond discussion and present the metrics used in the evaluation of results, generally related to the realism and consistency of the images produced. The metrics include subjective evaluation by people (Section 8.2) and/or objective evaluation comparing the synthesized image with the expected results (Section 8.1).

Another critical aspect of the area analyzed in this SR refers to facial expression image databases. These databases generally supply images with diverse facial expressions considering different subjects used as the basis for the synthesis of the desired facial expression. Some important variations are the different intensities of the facial expression and lighting conditions, among others. The relative completeness of these aspects renders the expressions database more attractive for synthesizing new images. A more detailed assessment can be found in Section 9.

Supplementary material is also available to provide an overview of the techniques presented in the articles in each of the stages analyzed. This material allows a complete view of each article, specifying which approaches were used in each stage. Further details of these approaches are found in Sections 4 to 8.

4 PREPROCESSING

Before applying techniques to obtain the synthetic image itself, many studies perform preprocessing stages that do not depend on the approach of synthetization adopted a posteriori. These stages can be divided into extraction of face landmarks, alignment of the faces, and selection of source images, as detailed in the sections that follow.

4.1 Facial Landmarks Identification

The first stage of preprocessing is the identification of facial landmarks. Landmarks are used to locate and define the face and facial components. The facial landmarks were labeled manually in 38% (20) of the studies, semi-automatically in 8% (4), automatically in 17% (9), and withdrawn from the facial expressions database in 4% (2) of the studies. Out of the remaining studies, 19% (10) did not specify and 15% (8) did not extract landmarks.

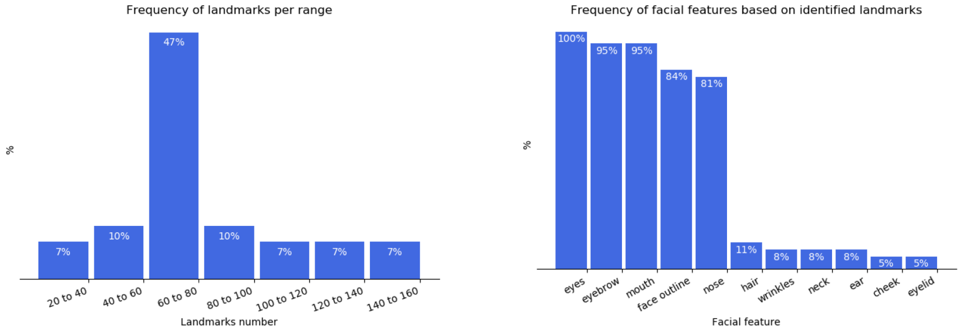
Manual identification is done by the database compiler using the source and target facial expressions. This approach is often used because of the possible inaccuracy of automatic identification or because landmark identification is not the focus of the study. Its advantage is that it ensures better results in identifying the points of the face and, consequently, better results in synthesis. However, its disadvantage is that it prevents system automation.

One way to facilitate the process of facial landmark identification is the semi-automatic approach. Some points of the face are previously selected manually and used as a basis for automatically identifying other points. The advantage of this approach is that it efficiently locates points in relation to the manual approach, especially when there is a large number of source images. Studies that adopted this approach located certain points manually and used feature templates (Li et al. 2007) or a Multi-Level Active Appearance Model (Lee et al. 2007, 2009) to locate the remaining points. Another semi-automatic approach was presented in Li et al. (2014), in which the internal points of the face were located automatically with an Active Appearance Model while the outermost points of the face were identified manually.

Although it facilitates the identification of facial landmarks, the semi-automatic approach still requires human interaction as part of the location process. Automatic approaches do not present this problem, but they can cause inaccuracies in processing. The automatic approaches adopted in the articles included Active Shape Model (ASM) (He et al. 2008; Xie et al. 2017, 2015), Active Appearance Model (AAM) (Xie et al. 2017, 2015, 2018), Local Texture Classifiers (Jia et al. 2010), Constrained Local Model (CLM) (Impett et al. 2014; Sabzevari et al. 2010), and Ensemble of Regression Trees (Wei et al. 2016). In (Fujishiro et al. 2009), the use of an automatic approach is mentioned but without presenting the method employed. In Xie et al. (2017, 2015), more than one technique is used to locate points because, according to the authors, ASM finds a greater number of points and AAM offers greater accuracy. A more complete overview of automatic detection techniques for facial landmarks can be found in Wang et al. (2017).

Xie et al. (2018) perform two additional tasks to improve landmarks identification: (i) obtaining a greater number of landmarks, and (ii) increasing the accuracy of their positions. The points are expanded through Procrustes transformations. Then the landmarks positions are finely matched and adjusted with a local optimization performed using B-splines and edge detection.

Another way to obtain information on facial landmarks identification is to use the facial expressions database. The advantage of this approach is that the labeling of landmarks is validated by the original study of the image database. In Macedo et al. (2006), the FGNET Database with Facial Expressions and Emotions from the Technical University of Munich (Wallhoff 2006) was used, whereas in Mohammadian et al. (2016), the database used was the Extended Cohn-Kanade Dataset (CK+) (Lucey et al. 2010).



(a) Number of facial landmarks identified in the images

(b) Landmark identification per facial feature

Fig. 4. Percentage of landmarks in the articles included.

Finally, some studies did not use facial landmarks to synthesize the new image. These studies directly manipulate the image without identifying points. In Kouzani (1999) and Wu et al. (2010), the authors directly use the pixels that compose the image, and in Chao and Zhiyong (2008) and Tay et al. (2009), facial regions that will be deformed to synthesize facial expression are presented by the authors.

A critical issue to be noted in relation to facial landmark identification is the number of facial landmarks detected. Figure 4(a) shows that most studies used between 60 and 80 points. However, there was great variation in the number, with a minimum of 26 and a maximum of 150 points. This variation occurs because the studies manipulate different parts of the face. For example, while some studies manipulate only the eyes and mouth region (Li et al. 2007), others also consider the regions of the eyebrows, eyes, mouth, nose, face outline, neck, hair, and ear (Li et al. 2014, 2012).

Despite the variation in the regions of landmark identification, Figure 4(b) verifies that most studies used the regions of the eyes, eyebrows, mouth, face outline, and nose. While those regions are used in more than 80% of the studies, the other regions of the face were considered by fewer than 12% of studies. These sparsely used regions are more difficult to detect (wrinkles and eyelids) or less important to the facial expression (hair, neck, and ear).

Landmark locating is not the only way to establish a correspondence between facial components among different faces. Another automatic approach is based on dense correspondences between two images. A dense correspondence between facial images is performed in Yu et al. (2018) through a Convolutional Neural Network (CNN) Flow. Further details are described in Section 6.

4.2 Facial Alignment

Facial images may come in a variety of sizes. Similarly, the placement of facial components can be in different positions in the source and in the target images. Faces should have the same dimensions if the components of source faces are to better suit target faces. To that end, the neutral and expressive source faces are aligned to the neutral target face. Another way is to align the source and target faces to an average face.

The alignment methods found to accomplish this task were affine transformation (15%; 8), Procrustes transformation (4%; 2), a combination of Procrustes analysis with affine transformation (6%; 3), pixel-based correspondence (2%; 1), Thin Plate Spline (TPS) (2%; 1), and normalization with Facial Animation Parameters Unit (FAPU) (2%; 1). Moreover, 21% (11) of the studies mention performing alignment without specifying the technique. Finally, 51% (27) of the studies did not specify if any alignment is performed.

Alignment often requires a previous step consisting in dividing the image into smaller pieces to allow the alignment of each facial component. In these cases, 21% (11) of the studies stated that triangulation was performed to accomplish the task. Regarding the studies that performed triangulation, more than half used Delaunay triangulation (Ghent and McDonald 2005; Impett et al. 2014; Li et al. 2014; Wei et al. 2016; Zhang et al. 2014; Zhang and Wei 2012).

4.3 Source Image Selection

This stage assumes that a database is already defined with neutral and expressive images that will be used to generate the target image. In this stage, the authors select which source images of the database will be selected for synthesizing the desired facial expression. In 66% (35) of the cases, all the images of the database are used to build the model. In general, the studies that adopted this approach are those that used machine learning and blend shape in the synthesis stages. About 28% (15) of the studies did not specify how the source images were chosen. The remaining studies (9%; 5) used similarity measures to choose images.

The approaches that use all the images available for constructing the model try to extract more information from this dataset. However, when there are no faces similar to the target face in the image database, the synthesized facial expression tends to be close to the average, thus losing the identity traits of the target face (Zhang and Wei 2012).

The studies that used similarity measures chose the source images by considering the following aspects:

Structure-based-similarity: The idea here is to find a similar source face using information on face structure or on a facial component. This approach assumes that subjects with similar neutral faces display similar expressive faces (Xiong et al. 2010, 2007b). In Xiong et al. (2010, 2007b), the most similar eye can be found by using a Euclidean distance-based structure similarity; however, this study is restricted to the synthesis of facial expression around the eye. In Mohammadian et al. (2016), the structure of face is considered as a whole and the similar face is found by means of the five nearest neighbors.

Expression-based-similarity: Facial expression databases are searched for a facial expression that presents the desired emotion (e.g., happiness, sadness, etc.). Then the application chooses the closest facial expression by means of facial feature motion difference computed using optical flow (Li et al. 2012) or with Mahalanobis distance and Information-Theoretic Metrics Learning (ITML) (Li et al. 2014). This image is used along with the source image to synthesize facial expression, as presented in Section 7.4.

5 DEFORMATION- AND MAPPING-BASED SYNTHESIS

Deformation and mapping is the first approach to be described for face image synthesis. It is used in 45% (24) of the articles in our survey. For the synthesis of new facial expressions, a neutral source face image, an expressive source face image, and a neutral target face image are usually required. In this category, there are also techniques that use only the expressive face as a source (disregarding the need for a neutral source face). This approach can be accomplished with the following steps:

(1) dividing the source and target faces into smaller parts;
(2) obtaining the new format for each division of the neutral target face by adding to it the difference between expressive and neutral source faces;
(3) deforming each division of the neutral target face to the new format;
(4) mapping the lighting differences on to the deformed face.

5.1 Deformation

In the first step of the deformation and mapping approach, deformation occurs, consisting of the movement of the control points that indicate an emotion in the facial expression. The face can be fully deformed (global deformation) or by regions (local deformation). None of the studies analyzed used global deformation as part of their approach. Only the study by Zhang et al. (2014) presented this type of deformation, but for demonstration purposes only. Therefore, most of the studies performed the division into smaller parts to be deformed as a first step. There are several techniques to perform deformation: of the studies included and that adopted the deformation and mapping approach, 48% (11) used triangulation and warping, 26% (6) used spline warping, 4% (1) used facial deformation tables with quadratic deformation, 4% (1) used a deformation based on edge direction preservation, and 17% (4) did not specify the deformation approach.

The deformation techniques mentioned show common and different points between them, namely:

Triangulation and warping: The face is divided into triangles that are fit into the new computed format. Delaunay triangulation was used in 91% (10) of the cases; the remaining studies did not specify the triangulation technique. Geometric-controlled image warping (Song et al. 2006; Xie et al. 2015; Zhang et al. 2012, 2014), affine transformation (Impett et al. 2014; Li et al. 2014, 2012; Wei et al. 2016; Zhang and Wei 2012), affine and smooth (Su et al. 2011), and bilinear Coons patch image warping (Leung et al. 1996) were used for warping.

Spline warping: This is the interpolation technique of a given surface based on spline curves. This technique is used with two-spline mesh warping (Lee et al. 2007, 2009; Sabzevari et al. 2010) and TPS warping (Chao and Zhiyong 2008; Jia et al. 2010; Li et al. 2007).

Facial deformation tables with quadratic deformation: In this technique, the face is divided according to the regions of the muscles. Then, from information about muscles movement in each region, the face is deformed by means of quadratic models (Tay et al. 2009).

Deformation based on edge direction preservation: Xie et al. (2018) deformed the triangular mesh using a nonlinear least squares optimization that preserves facial feature geometry details by directional vectors along multiple boundary edges. This approach concerns not only preservation of the exterior shape (face profile), but also of interior component (facial features) shapes.

Deformation allows changing the image based on changes of facial components formats. That is, the format of the facial components of the target face is adjusted according to the changes identified in the source face.

5.2 Mapping

Mapping is the second step of the deformation and mapping approach. In this step, lighting information related to facial expression is mapped to the target face. The idea is to transfer the lighting changes that occur between the expressive and neutral source images. To that end, 48% (11) of the studies that adopted the deformation and mapping approach used Expression Ratio Image (ERI), 15% (3) used lighting differences, 10% (2) used blend ratio, 5% (1) used Helmholtz-Hodge Decomposition (HHD), and 5% (1) used frequency analysis. Nineteen percent (6) did not map lighting information.

The techniques used in this step differ depending on technique:

ERI: ERI produces an image that captures lighting changes as the ratio between the neutral face source and the expressive face source. The main idea of ERI is to use lighting variations to describe facial changes (Li et al. 2014). The first authors to introduce this concept were Liu et al. (2001). It is the most widely used technique for transferring lighting information between facial expressions; however, it can present artifacts due misalignment in the warping processing. This occurs because ERI is sensitive to variations in pixel intensity (Song et al. 2006; Xie et al. 2017). A way to improve the accuracy of ERI is to increase the correlation between the pixels of the source image and those of the target image. This can be done by using a curves model (Xiong et al. 2010, 2007b). Another way is to highlight the areas of lighting transfer based on information about muscle movement (Zhang et al. 2012, 2014). Finally, the difference in angle can also be considered to improve ERI's accuracy (Lee et al. 2007, 2009).

Lighting differences: The lighting difference technique avoids the problem of correspondence between the pixels that is present in the ERI technique. Lighting differences approaches include the difference between source images (Xie et al. 2017, 2015, 2018) and the difference between mean values of the neutral target image and neutral source image (illumination bias) (Zhang and Wei 2012).

HHD: This is another option to minimize the artifacts of ERI (Song et al. 2006). HHD consists of a lighting transfer technique using the concepts of vector field decomposition. The transformation matrix is built from components of divergence and rotation obtained with HHD (Song et al. 2006).

Frequency analysis: Frequency domain analysis is used to better capture information about the facial expression. The study by Wei et al. (2016) divides the image into four frequency bands for transferring the details of facial expressions.

The mapping of lighting information allows one to transfer the changes in lighting from the source images to the target images. This transfer especially maps the finer details, such as wrinkles and creases. This makes the synthesized images more consistent with actual images.

6 MACHINE LEARNING-BASED SYNTHESIS

Machine learning is the second approach presented for facial image synthesis. It is used in 66% (35) of the articles included in our survey. In those studies, the desired facial expression is learned based on multiple instances of that expression. New images are synthesized using a model built from the expressive facial features patterns found in the dataset of the source images. This process can be conducted through a supervised or unsupervised machine learning approach. Ninety-seven percent (30) of the machine learning studies used supervised methods (Section 6.3.1), and only one study used unsupervised learning (Section 6.3.2).

Three steps are performed: (i) feature extraction, (ii) dimensionality reduction, and (iii) synthesis. For step (i), Section 6.1 describes which features are considered by these methods. For step (ii), all the studies reduced data dimensionality using methods based on feature fusion. The third step, synthesis, is performed by two alternative approaches: (i) The new image is reconstructed using the features decomposition performed in the second step (Section 6.2), or (ii) a model is trained using the fused features and is used to synthesize the new images (Sections 6.3.1 and 6.3.2).

6.1 Facial Features Extracted

In this section, we consider which features are extracted from the example images and used for learning the expression models. The studies considered features extracted from the images based on their shape (83%; 29), texture (86%; 30), or illumination (11%; 4).

Shape: Shape features are based on changes in facial structures. For this, 74% (26) of the studies used the positions of facial landmarks. The others used an approach to define the face in small patches (He et al. 2008; Mohammed et al. 2009; Wang and Wang 2008). To discover changes to the shapes of facial components is the first step in changing the shapes of those components. This change can be combined with learning the texture or lighting or with deformation approaches (Section 5.1).

Texture: Some studies consider the positions and relationships between the pixels of the image as features to be learned. These features can be combined with changes to the shape. There are also studies that perform postprocessing by means of mapping illumination information (Section 5.2).

Illumination: There are many ways to consider the color and light information of an image to build the model to be learned: ERI (Du and Lin 2003; Mima et al. 2011; Zhu et al. 2004), illumination bias (Wei et al. 2016; Zhang and Wei 2012), and lighting differences (Xie et al. 2017) are the main ones found in the articles included in our survey. These are part of the mapping approaches to illumination information found in Section 5.2. These features were combined with shape and deformation approaches (Section 5.1).

Combinations: Each extracted feature can be combined with another for more information. The combination of shape and texture was found in 60% (21) of the machine learning studies, commonly using landmarks and the vector of pixels. On the other hand, the combination of shape and illumination was made with landmarks and ERI in 9% (3). There are two ways to make this combination: considering each feature separately or together for constructing the model.

Other features: Some approaches, such as deep neural networks, have an encoder that performs facial features extraction and dimensionality reduction of the image inside the model. These approaches are described in Section 6.2.

6.2 Dimensionality Reduction

Only feature fusion methods were used by the studies for dimensionality reduction. Fusion of features represents a class of methods for dimensionality reduction that combines original features in new features. These new features are usually sorted by some relevance criterion and only the most relevant are used in the model learning. In this process, the number of new features used is reduced in comparison to the original feature set.

The feature fusion techniques can also be used to build a model from a linear combination of the transformed features, and, therefore, it does not perform a subsequent learning step. The main idea is to decompose the images into two subspaces of features: a subspace of expression features and a subspace of identity features. These subspaces are used to reconstruct the new image. This was carried out using different techniques.

Principal Component Analysis (PCA): PCA is used to identify facial changes with greater variability. The desired facial expression can be synthesized by the weighted projection of the basis images (eigenfaces) (Du and Lin 2002; Kouzani 1999; Mohammadian et al. 2016), or PCA output is used in conjunction with learning algorithms. PCA was applied to landmarks (Ghent and McDonald 2005), landmarks with pixel vectors (Agarwal et al. 2012; He et al. 2008; Li et al. 2007; Seo and Chen 2012; Wang and Ahuja 2003; Zhou and Lin 2005), and the ERI logarithm (Du and Lin 2003).

PCA Variations: Some studies used alternative versions of PCA in their approaches. These variations were Quadtree PCA (QPCA) (Kouzani 1999), Multilinear PCA (MPCA) (Macedo et al. 2006), PCA applied to each facial expression separately (Agarwal et al. 2012), and, in Seo and Chen (2012), PCA applied to training images and residual error.

Statistical Shape and Appearance Models: This technique consist of statistical models built to represent the shape and appearance of the face in following the stages: facial landmarks identification (Section 4.1), facial alignment (Section 4.2), and identification of the main facial variations using PCA. The models used were ASM (Xiong et al. 2007a) and AAM (Abboud and Davoine 2004, 2005; Abboud et al. 2004; Macedo et al. 2006; Mohammadian et al. 2016; Tan et al. 2009; Xie et al. 2017; Xiong et al. 2009; Zhang and Wei 2012).

Singular Value Decomposition (SVD): SVD is used to decompose the face into the subspaces "face identity" and "facial expression." Thus, it is possible to reconstruct a certain identity with another expression. In addition to the use of Support Vector Regression (SVR) (Lee et al. 2007), some variations were identified, such as HOSVD (Lee and Elgammal 2006; Macedo et al. 2006; Tan et al. 2009; Wang and Ahuja 2003; Wei et al. 2016; Zhang and Wei 2012), kernel-based bilinear factorization (Zhou and Lin 2005), and bilinear factorization symmetric and asymmetric models (Abboud and Davoine 2004, 2005; Mohammed et al. 2009). As in PCA, some studies used decomposition as input for learning (Lee et al. 2007; Wei et al. 2016; Zhang and Wei 2012; Zhou and Lin 2005). Other studies reconstructed the decomposed faces. Finally, most studies applied SVD in landmarks and image vectors, but it was also applied to image patches (Mohammed et al. 2009).

Supervised Locality Preserving Projections (SLPP): SLPP aims to preserve the characteristics of the facial expressions locally by means of a linear approximation of the Laplacian eigenmap. In Wang and Wang (2008), the face is divided into patches and their weighted reconstruction is made to synthesize the new facial expression.

Face encoding: The main goal of the encoder in a neural network is features extraction and dimensionality reduction. In the case of facial expression synthesis, the encoder extracts components of two subspaces: identity and facial expression. In Moghadam and Seyyedsalehi (2018), the encoder also considers the expression intensity. Zhou and Shi (2017) use a CNN to encode a 32×32 colored image of the face into a feature vector of the latent space. Moghadam and Seyyedsalehi (2018) encode a face in the Dynamic Deep Bottleneck Neural Networks (DDBNN) model. Yu et al. (2018) present a more general approach to match regions between images, thus the features are the semantic correspondence.

6.3 Model Building

6.3.1 Supervised Learning. The supervised methods used for model training were regression, Support Vector Machine (SVM), neural networks, and deep neural networks, detailed below.

Regression: Some of the studies used linear regressions, multiple linear regression, polynomials, and SVR. In the case of linear regressions, the inputs were PCA (He et al. 2008), AAM [Abboud, Davoine, and Dang 2004], and illumination differences (Abboud et al. 2004) (Section 5). For the multiple linear regressions (Mima et al. 2011), features resulting from PCA were used. For the polynomial (quadratic), both the image and ERI were used (Du and Lin 2003; Xiong et al. 2007a). Finally, SVR used the features resulting from PCA to build the model (Zhu et al. 2004).

SVM: SVMs whose kernel functions are Gaussian RBFs (Agarwal et al. 2012; Fujishiro et al. 2009; Wei et al. 2016; Xie et al. 2017; Zhang and Wei 2012) and inverse multiquadric RBFs (Lee et al. 2007) were used. The input for these SVMs are features resulting from PCA (Agarwal et al. 2012; Xie et al. 2017), SVD (Lee et al. 2007), AAM with SVD (Wei et al. 2016; Zhang and Wei 2012), or wrinkles (Xie et al. 2017). The SVD calculates only the movement of the facial landmarks; then, the face is deformed using deformation and mapping approaches (Section 5.1).

Neural Networks: Neural networks are used in two different ways. The first concerns the RBF network (Ghent and McDonald 2005; Leung et al. 1996), and the second is discussed in the deep neural networks description below. An RBF network is a type of neural network that uses Gaussian functions as the activation function. The neural network's input are features resulting from PCA (Ghent and McDonald 2005) or the vector of facial landmarks. In the case of facial landmarks, the patterns of movement are learned, then the target image is deformed and the mapping of illumination information is made, as shown in Section 5.

Deep Neural Networks: Three types of deep neural networks were found: CNN Flow, DDBNN, and Deep Belief Network. The CNN Flow (Yu et al. 2018) hierarchically matches the regions between two images in a dense correspondence. It uses a CNN feature pyramid to establish the correspondence between the semantic features. Then it generates a coarse target neutral image according to its correspondence to the source image. The DDBNN (Moghadam and Seyyedsalehi 2018) synthesizes the new image from the identity manifold, facial expression, and the intensity of expression (see Section 6.2). The model uses recurrent networks to generate images with different levels of intensity. Finally, the Deep Belief Network (Sabzevari et al. 2010) defines the target expressive landmark positions by extracting variations in facial expressions. This network is trained using landmarks and FACS labels. Then the face is deformed according to the learned landmarks using warping (see Section 5.1).

6.3.2 Unsupervised Learning. Seventy-five percent (3) of the publications concerning unsupervised learning use Generative Adversarial Networks (GAN). Only one study used cluster and correlation to recognize the movement patterns of a specific facial expression in order to make a corresponding deformation in the target face.

Cluster and correlation: In Jia et al. (2010), the facial landmarks are divided into face regions according to FAPs (MPEG-4 facial animation parameters). The FAPs are then normalized with FAPU in order to be clustered using a hierarchical clustering tree. Then a Pearson correlation coefficient identifies the movement patterns used for deforming the target face through a thin plate spline (see Section 5).

Generative Adversarial Networks: Zhou and Shi (2017) train a Conditional Difference Adversarial Autoencoder (CDAAE) from the encoded faces (Section 6.2) to learn the difference between the source and target images. Then the decoder synthesizes the expressive face. Choi et al. (2017) train the model called StarGAN to learn image-to-image translation across multiple domains. StarGAN uses a single generator to manipulate facial attributes such as expression, age, and gender, among others.

7 SYNTHESIS WITH OTHER APPROACHES

Other approaches, found in fewer of the included studies, presented different solutions to the problem. The approaches of blend shape (Section 7.1), muscle simulation (Section 7.2), and cloud modeling (Section 7.3), and others (Section 7.4) were used.

7.1 Blend Shape

This approach is divided into two steps: the division of the face into smaller parts and the blending of the regions containing the facial expression on the source images. The division is made into subregions defined by the study authors and considering facial components. The blend of images is through interpolation with pixel-wise color blending and merging through weight subdivisions map (Zhang et al. 2003, 2006).

According to the authors, the approach can realistically synthesize the desired facial expression. However, in studies presented by Zhang et al. (2003, 2006), some photographs of the same person to be synthesized (target image) are used as source images. Therefore, adjustments would be required for the operation when using source faces different from the target face.

7.2 Muscles Simulation

Some articles used information from muscle movement as a way to improve a specific aspect of the synthesis of facial expressions. This information was used in the process of deforming the face and in the mapping of features, both presented in Section 5. That is, the information obtained from the model of muscles is used as a way to improve some deformation and illumination mapping approaches.

In Tay et al. (2009), the information about the muscles is used to define the regions, direction of deformations, and intensity of quadratic deformations. A deformation facial table defines the regions to be deformed, as well as the direction of deformation.

Zhang et al. (2012, 2014) highlight the information about ERI illumination transfer in areas where there are muscle movements. Thus, there is an attempt made to decrease the amount of information not related to facial expression present in the ERI generated by source images. The information about the muscles restricts the regions for the ERI mapping.

7.3 Cloud Model

In this approach, synthesis is based on a model of transformation of uncertainty between the quantitative model (based on the image) and the qualitative concept (facial expression). The quantitative model is formed by numerical characteristics: expectation, standard variance, and hyper-entropy (the measure of uncertainty of the standard variance) (Wang et al. 2012).

The image is synthesized by means of forward cloud generator and backward cloud generator (Wu et al. 2010). The forward cloud generator transforms each image into its respective numerical characteristics. Finally, the numerical characteristics of each image are transformed into a cloud with the backward cloud generator. The reconstruction of facial expression is accomplished by the weighted combination of these clouds (Wu et al. 2010).

7.4 Similar Facial Expression

This approach combines two images to generate the expressive target face. The first image results from the deformation and mapping process (Section 5). The second image is the result of the search for a person's facial expression most similar to the expressive source face, as shown in Section 4.

Both images—found and synthesized by deformation and mapping—contain expressive faces close to the desired facial expression. Therefore, the end result is achieved by merging these images. This refinement of the facial expression aims at obtaining a more realistic result.

8 EVALUATION OF THE RESULTS

A crucial issue in the synthesis of images is how to check if the result (the resulting images) is appropriate. The evaluation metrics found in the studies included in our survey were divided into objective and subjective approaches. Objective evaluation metrics are those that evaluate the results quantitatively, considering only the aspects of images, mainly using computational techniques. Subjective metrics are those in which people evaluate the performance of the system.

Objective metrics (Section 8.1) were observed in 26% (14) of the studies and subjective metrics (Section 8.1), in 19% (10). The remaining 55% (29) of the studies only discussed the synthesized images (Section 8.3) without presenting a subjective or objective evaluation. These aspects are presented in Section 8.3.

8.1 Objective Metrics

For an objective evaluation, it is necessary to analyze the objects evaluated and the techniques used to perform the task. The analysis is carried out according to a metric that does not depend on a person's perception. The most widely used object for evaluation was the synthesized image (90% of the objective metrics). Facial landmarks were the only other object evaluated.

The synthesized images were compared with other images, such as ground truth images, images synthesized by other studies, and expressive source images. The comparison with ground truth images allows verifying if the synthesized image meets expectations. A comparison with synthesized images made by other studies allows an analysis of differences between approaches. Finally, the comparison with the source image allows verifying the consistency of the synthesized images. The metrics found can be divided into two categories: those that use a reference image and those that do not use a reference image (called no-reference). While the former evaluates the difference between two images, the latter uses statistical aspects of how the images should look or about the possible distortions (Wang and Bovik 2006).

We found the following metrics related to the first category:

Mean Square Error (MSE): This is the most widely used technique to compare images. In Agarwal et al. (2012), Seo and Chen (2012), Xiong et al. (2009), and Zhou and Lin (2005), MSE was used to compare synthesized images with the ground truth, images synthesized by other approaches, and with source images. In addition, there were also variations of the MSE: ratio Gradient Minimum Square Error (GMSE) and average Peak Signal to Noise Ratio (PSNR). The GMSE was used to compare the synthesized image with the target neutral image to verify its quality and identity preservation (Wang and Ahuja 2003). The PSNR is a way to evaluate noise interference in the synthesized image (Xiong et al. 2009). The relation of the source images to the synthesized images was evaluated using this method. Xiong et al. (2009) also compare the approach proposed by their study with other approaches from the literature.

Correlation: Correlation measures the dependency between two images to determine their statistical relationship. Correlation coefficient (Ghent and McDonald 2005), Pearson product-moment correlation (Du and Lin 2003), and correlation of lighting differences (Xie et al. 2017, 2015, 2018) were used. The studies correlated the synthesized images with the ground truth, target neutral, and synthesized images with source images.

Robustness: Xie et al. (2018) measured the ability of their algorithm to withstand noise. They calculate the minimum sum of the 1,000 largest differences between ∇source and ∇target images, where ∇ is the difference between expressive and neutral images.

Structural similarity: The Structural SIMilarity (SSIM) index aims to evaluate the quality of an image using structural similarity in the space domain (Wang and Bovik 2006). Moghadam and Seyyedsalehi (2018) use the SSIM to compare the synthesized image with the target neutral image.

Consistency of landmarks: Abboud et al. (2004) evaluate the distances of facial landmarks through the Box-Whisker Diagrams metrics. These metrics verify whether the measures found for the synthesized images are within an acceptable range. Verifying the distances for the synthesis allows one to evaluate the consistency of the shape.

Facial expression recognition: Agarwal et al. (2012) and Choi et al. (2017) trained a facial expression recognition classifier to automatically evaluate whether a synthesized image properly displays the desired facial expression. Choi et al. (2017) check the accuracy by comparing the results with other approaches as well as with real images.

The only metric found in the second category was Q-value. Q-value is a spatial domain method in which the image is partitioned into blocks to evaluate the effects of blocking and blurring on a given image (Wang and Bovik 2006). Thus, it is possible to evaluate local blockiness, such as those caused by JPEG compression (Wang et al. 2002).

8.2 Subjective Metrics

In studies using a subjective evaluation of results, the following aspects were verified: number of people who performed the evaluation, characteristics of the evaluators, evaluation task, and consistency mechanism.

The number of people who evaluated refers to the number of volunteers or paid users who conducted the evaluation tasks. This number varied from 8 to 112 people, most commonly around 30 evaluators.

The characteristics of those who evaluated results help to establish patterns according to users' profiles. More than half of the studies that use subjective evaluations did not mention the characteristics of the people who evaluated the synthesized images. The other studies raised characteristics related to age group, gender, educational attainment, the existence of a visual impairment, and knowledge about the cloning of facial expressions.

In general, in the subjective evaluation, the synthesized images are subjected to evaluation tasks to verify the performance of the proposed approach. The tasks applied were the perception of realism of the images generated, the identification of facial expression, and the preservation of identity. Listed here are the tasks for each of these evaluations.

Image realism: The evaluators give a grade from 1 to 5 for each synthesized image (Li et al. 2014; Wang and Wang 2008; Xie et al. 2017; Zhang et al. 2012, 2014) or 5 to 10 (Xie et al. 2018). The goal is to evaluate whether there are elements that make the images different from a real image. It is thus possible to verify the consistency of synthesized images.

Facial expression identification: This task classifies what facial expressions are shown in the image (Choi et al. 2017; Li et al. 2007; Tay et al. 2009; Wang and Ahuja 2003; Wang and Wang 2008). For this, the evaluator selects the facial expression that best fits into the group of facial expressions available, which verifies if the synthesized facial expression has the desired facial expression elements.

Pearson identity preservation: In Wang and Wang (2008), this task is carried out in two supplementary ways: person verification and double-blind face recognition. In person verification, the synthesized image and ground truth are placed side by side and the person who is evaluating checks whether the images belong to the same person. Double-blind face recognition uses a group of synthesized images and a group of images from the target person. The images of the first group should be related to the second group. Zhou and Shi (2017) asked users to verify one of three images that best corresponds to the same person's ground truth image. The images were synthesized by three different approaches and presented in random order. Evaluators could select one or no image. These tasks allow evaluators to check if a person's specific traits have been lost during the synthesis process.

Some studies also use mechanisms to verify the consistency of the evaluators' responses. In addition to the synthesized images, other images are evaluated to ensure the consistency of the responses provided. The images included in the tests are ground truth (Wang and Wang 2008) or synthesized by means of other approaches (Zhang et al. 2012, 2014). The first one allows checking that the evaluator responds in a manner befitting a real image, while the second allows comparing the results with those of other studies.

Finally, a statistical analysis can be used to validate the results of the subjective evaluation. To achieve this, Yu et al. (2018) applied a Wilcoxon rank sum test to compare the results of their approach over other works. The applied test evaluates the results of the verification task on the realism of the images.

8.3 Results Evaluation by the Authors

This section describes aspects of the synthesized images that the authors discussed in their results. The most common aspects described by the analyzed studies are the provision of varied images (60%; 32), comparison with images synthesized by other approaches (36%; 19), wrinkles (19%; 10), and a comparison with ground truth images (17%; 9).

One of the aspects highlighted by the authors and analyzed in their discussions is the need to present several aspects of variation in the faces synthesized. The studies aim to generate different facial expressions, different faces/people, variations in the intensity of facial expressions, wrinkles, mixed facial expressions, transitions between two facial expressions, identity preservation, and exaggerated facial expressions. The variation of aspects present in the synthesized images allows one to analyze the approach's ability to synthesize images that are useful for different contexts.

Another way to show the operation of the approach is through a comparison of results. Comparisons were made between synthesized images and those generated by other studies or ground truth images. This allows researchers to check if the synthesized images are as suitable as the images used as reference. This approach was discussed in Section 8.2, under subjective evaluations with users. However, several studies just discuss this aspect without using subjective metrics.

Finally, there are conditions that influence the process of synthesis of facial expressions. There were reported tests considering different illumination conditions, improvement in the precision of the detection of facial landmarks, different ways to find similar images in the image database, blurriness, invalid settings of facial expression, occlusion, people not present in the source image database, pose variations in the target face, synthesis in non-human faces, and runtime. Showing the behavior of the approach under these varied conditions allows researchers to verify if the approach works even in the most adverse conditions. It can also define the conditions required to improve a system's functioning.

These aspects were discussed as ways in which the studies described and evaluated their results. An adequate set of various aspects helps to check if the synthesis of the expressions occurred consistently in each perspective observed. Thus, these characteristics can evaluate the efficiency of the solution presented.

9 FACIAL EXPRESSION IMAGE DATABASES

The facial expressions image databases mentioned in the studies included in this SR contain the images used as a basis for synthesizing facial expressions (source images) and serve to evaluate the approaches proposed by the authors. In addition to the facial expression databases retrieved from the analyzed publications, we added two popular facial expression databases: FER2015 (Goodfellow et al. 2013) and Oulu-CASIA (Zhao et al. 2011) due their relevance to this article.

Most of the analyzed publications concerned CK (Kanade et al. 2000) (23%; 12), CK+ (Lucey et al. 2010) (19%; 10) and JAFFE (Lyons et al. 1998) (13%; 7) databases, whereas each of the remaining ones are used by one (2%) or two (4%) studies each. The complete list of these databases is found in Table 4. In addition to these, 13% (7) of the studies captured their own images. The remaining 26% (14) did not specify the acquisition of the images used. The sum of these percentages results in more than 100% because there are studies that used more than one image database. Note that the database CK+ (Lucey et al. 2010) is a continuation of CK (Kanade et al. 2000). Likewise, the CMU MultiPIE (Gross et al. 2010) is a continuation of CMU-PIE (Sim et al. 2003).

Table 4. Main Points of the Discussions (Section 8.3) Observed on the Facial Expressions Image Databases

Database	Facial expressions	Same neutral and expressive face	Diversity in wrinkles	Number of different faces	Total images	Intensity of facial expressions	Different lighting conditions	Number of studies
CK (Kanade et al. 2000)	Joy, surprise, sadness, disgust, anger, and fear	Yes	Yes	97	486 (sequences)	Yes	Yes	23% (12)
CK+ (Lucey et al. 2010)	486 action units including: joy, surprise, sadness, disgust, anger, and fear	Yes	Yes	123	593 (sequences)	Yes	Yes	19% (10)
JAFFE (Lyons et al. 1998)	Neutral, happiness, sadness, surprise, anger, disgust and fear	Yes	Yes	10	213 (images)	Yes	No	13% (7)
AIAR (Fu et al. 2003)	Neutral, exaggerated and non-primary expressions	Yes	Unspecified	10	300 (images)	Unspecified	No	4% (2)
RaFD (Langner et al. 2010)	Neutral, sadness, contempt, surprise, happiness, fear, anger, and disgust	Yes	Yes	67	8040 (images)	No	No	4% (2)
POFA (Ekman and Friesen 1975)	Neutral, happiness, anger, fear, surprise, sadness, disgust, and contempt	Yes	Yes	16	110 (images)	No	No	2% (1)
XM2VTS database (Messer et al. 1999)	Speech recordings	Yes	No	295	1.180 (images)	No	Yes	2% (1)
IMM face database (Nordstrøm et al. 2004)	Neutral, happy and arbitrary expression	Yes	No	40	240 (images)	No	Yes	2% (1)
FG-Net Facial Expressions and Emotions Database (Wallhoff 2004)	Neutral, happiness, sadness, surprise, anger, disgust and fear	Yes	Yes	18	399 (sequences)	Yes	Unspecified	2% (1)
CMU-PIE (Sim et al. 2003)	Neutral, smile, blink, and talk	Yes	Unspecified	68	40.000 (images)	Yes	Yes	2% (1)
CMU MultiPIE (Gross et al. 2010)	Neutral, smile, blink, and talk	Yes	Unspecified	337	750.000 (images)	Yes	Yes	2% (1)
DISFA (Mavadati et al. 2013)	FACS and pain intensity coding	Yes	Yes	27	4.845 (video frames)	Yes	No	2% (1)
FER2013 (Goodfellow et al. 2013)	neutral, happiness, sadness, surprise, anger, disgust and fear	No	Yes	Unspecified	35.887 (images)	No	No	-
Oulu-CASIA NIR-VIS (Zhao et al. 2011)	neutral, happiness, sadness, surprise, anger, disgust and fear	Yes	Yes	80	2.880 (sequences)	Yes	Yes	-

The discussion of results (Section 8.3) analyzes which aspects are important in the facial expression databases. The aspects most widely used by the studies are reported in Table 4. These aspects are essential to check if the approach can synthesize different facial expressions, compare images with the ground truth (same neutral and expressive face), produce variations of expressions, synthesize expressions in several faces (identity preservation), generate different intensities, and consider lighting conditions. In addition to the aspects shown in Table 4, other aspects were highlighted as necessary by a smaller number of studies.

(1) Facial expressions related to more than one emotion: present in CK+ (Lucey et al. 2010), CK (Kanade et al. 2000), and DISFA (Mavadati et al. 2013);
(2) The position of the landmarks to check the accuracy of the location of these points: present in CK+ (Lucey et al. 2010), CK (Kanade et al. 2000), IMM face database (Nordstrøm et al. 2004), FG-Net Facial Expressions and Emotions Database (Wallhoff 2004), and CMU MultiPIE (Gross et al. 2010);
(3) Ethnic diversity to find faces similar to the target face: present in CK+ (Lucey et al. 2010), CK (Kanade et al. 2000), CMU MultiPIE (Gross et al. 2010), DISFA (Mavadati et al. 2013), Oulu-CASIA NIR-VIS (Zhao et al. 2011), and FER2013 (Goodfellow et al. 2013);
(4) Spontaneous facial expressions: present in DISFA (Mavadati et al. 2013) and partially covered by CK+ (Lucey et al. 2010), CK (Kanade et al. 2000), and FER2013 (Goodfellow et al. 2013);
(5) Different poses to cover expressions in wild: present in the databases CK+ (Lucey et al. 2010), CK (Kanade et al. 2000), IMM face database (Nordstrøm et al. 2004), CMU MultiPIE (Gross et al. 2010), RaFD (Langner et al. 2010), and FER2013 (Goodfellow et al. 2013).

Finally, some aspects were not found in any of the databases. These aspects are non-human faces, blurriness, and occlusion. Such aspects were not often used to discuss results.

The databases that contain most of these aspects contributed to obtaining more accurate synthesized images and to obtaining better evaluation of results. Most aspects are covered by the databases, but sometimes only partially. For example, JAFFE (Lyons et al. 1998) and AIAR (Fu et al. 2003) only have ten different faces, while CMU MultiPIE (Gross et al. 2010) has 337. Most databases offer images of six universal emotions and the neutral facial expression. Additionally, the CK+ database (Lucey et al. 2010) offers face movements related to 486 action units defined by the FACS, representing more than six emotions, while CMU MultiPIE (Gross et al. 2010) and CMU-PIE (Sim et al. 2003) offer only images of smiling, blinking, and speech. Finally, the databases AIAR (Fu et al. 2003), CMU MultiPIE (Gross et al. 2010), and CMU-PIE (Sim et al. 2003) provide images related to speech movements. Although none of the databases has all aspects, the databases analyzed fulfill the necessary conditions for synthesizing facial expressions.

10 DISCUSSION

The following sections present our main considerations made from the analysis of the set of articles included in this SR. In addition to the analysis of article trends included in the SR, challenges are discussed as well as ways to overcome them, which may indicate research opportunities.

10.1 Preprocessing Methods

The preprocessing of images allows adjusting the source images to adapt them to the target image and/or to extract information about shape. However, some approaches do not require this step.

Table 5 shows the strengths and weaknesses of each preprocessing approach. The optimization attempts are based on choosing the closest facial expression and the most similar facial structure. While the first one searches an alternative target face, the second tries to obtain a more similar facial expression through a more similar neutral face.

10.1.1 Source Image Selection. The selection of the appropriate source images can optimize the synthesis result. However, only 9% (5) of the studies reviewed in this SR used some form of automatic selection of those images, as seen in Table 5.

Table 5. Strengths and Weaknesses in the Preprocessing Approaches by Stage

Stage	Approach	Number of studies	Strengths	Weaknesses
Source images selection	Expression-based-similarity	4% (2)	Target image has natural expressiveness before the synthesis.	Requires expressive images of the target face.
	Structure-based-similarity	6% (3)	Source face tends to have a facial expression more similar to the target face.	Requires faces structurally similar to the target face to guarantee the best results.
Landmarks identification	Manual	38% (20)	Guarantees good accuracy of points.	Difficult achievement for a large portion of the images.
	Facial expression database	4% (2)	Guarantees good accuracy of points.	It is restricted to the images of the databases that provide this information.
	Semi-automatic	8% (4)	Identification of points with less human intervention.	Depends on the accuracy of human identification and algorithm.
	Automatic	17% (9)	Easy location regardless of the number of images.	Depends on the accuracy of the detection algorithm.
Facial alignment	Image transformations	32% (17)	Allows working directly with the pixels.	Requires more processing.
	FAPU	4% (2)	Enables transformations that are not possible for the images.	Requires a separate approach to transform images.

To guarantee better results, these approaches require certain conditions in the facial expressions databases, as explained in Table 5. Images of expressive faces of the target person are often unavailable in the databases; therefore, the expression-based similarity approach (Section 4.3), can only be applied to some applications. In the case of structure-based similarity (Section 4.3), an image database with a variety of ethnicities, ages, and genders is required to produce a greater chance of finding a really similar face.

In the studies analyzed, only two types of similarity measures for image selection were found, both in relation to shape (landmarks) and considering the neutral face or facial expression. It can be a good idea to use other types of similarity measures based, for example, on pixel colors and the relationships between facial components. Another type of similarity considers each facial component separately and uses each source image found for the synthesis of a specific facial component. Ideally, a measure of similarity should be found that could guarantee a source face that works best with the target face regardless of the number of examples available.

None of the studies analyzed used more than one similarity measure for image selection. Therefore, another possibility of optimization is to combine several measures of similarity. For example, color information about the image combined with the structure of the facial component can be used to select the closest face. The composition of measures can guarantee a better result, especially if one of the measures is common to several images from the database. In this case, the use of additional measures could refine the search result.

It is also possible to select images without using a similarity measure. For example, it is possible to obtain the most expressive images set (representing an emotion with more intensity) as an image source. Hence, the source images would always be those with greater expressiveness. The challenge of this approach is to label the images of the database to identify them when necessary. This labeling could be performed either manually or automated.

10.1.2 Facial Landmark Identification. The identification of facial landmarks occurred in 87% of the studies analyzed, indicating that this information is important for synthesizing the desired expression. There are few alternative representations that perform the same function.

Not all studies consider that this aspect is part of the scope of the synthesis of facial expressions. This is verified in the difference between the number of studies that use the location of landmarks (Table 5) and the number of studies that specify the way they were acquired (Table 3). However, it is an almost essential step considering that most of the studies analyzed use this information to perform deformation and to learn during the steps of facial image synthesis.

Most of the studies made this identification manually, as shown in Table 5. This is a way to ensure the accuracy of facial landmarks, eliminating the concern for implementing high-accuracy systems for detecting these points. Albeit accurate, this approach is time-consuming and the results can be variable depending on the knowledge and the fatigue level of those who perform the task. Therefore, there are also a large number of automated systems for this task, such as DLIB[6] (Kazemi and Sullivan 2014), OpenFace[7] (Baltrusaitis et al. 2013), clmtrackr[8] (Saragih et al. 2011b), and Menpo[9] (Alabort-i Medina et al. 2014). The manual approach does not allow full automation of facial expression synthesis. In some types of synthesis applications, the automatic processing of a new image related to the neutral target face is necessary. For this type of application, automation of this step is also required.

The more automated the identification of landmarks is, the faster a large number of faces is obtained. However, this is one more step to be produced and it depends on the accuracy of detection, as shown in Table 5. The inaccurate identification of these points can lead to inconsistencies in synthesized images. Studies attempting to improve the accuracy of the identified landmarks (see Section 4.1) show that the points located by automatic approaches do not always have enough precision.

The realism of a facial expression may be related to the detailing of its facial components. The main components of the face (eyes, mouth, eyebrows, nose, shape of the face) were used by most studies analyzed. However, some components, such as the cheeks and wrinkles, have received little attention, as can be seen in Figure 4(a). Therefore, most studies do not carefully address these regions, which would be interesting for a more realistic result. Disregarding such points may be due to the absence of their representation in classical landmarks related to the facial expression of emotions mentioned in the literature.

Another factor that influences the detailing of synthesized facial expressions is the number of facial landmarks. A good demarcation of the components of the face depends on that number. Most studies use between 60 and 80 points, as shown in Figure 4(b). This quantity is good enough for the most widely used regions (eyes, mouth, eyebrows, nose, shape of the face), but not for others (wrinkles, eyelids, cheeks, neck, ears, and hair). A study even expands the number of points after identification. This leads to the conclusion that a larger number of points allows a more precise delimitation of each component, which can result in a better outcome of the synthesis. For example, the detection of landmarks related to wrinkles allows a demarcation of regions where they appear; thus, it is possible to avoid the transfer of aspects unrelated to the desired expression (Xie et al. 2017, 2015). For example, the pose could vary the lighting conditions of a particular region of the face, thus modifying the mapped regions according to the approaches presented in Section 5.2. However, ready-made libraries may be limited in this aspect. This could lead authors to

[6] http://dlib.net/.

[7] http://www.cl.cam.ac.uk/research/rainbow/projects/openface/.

[8] https://github.com/auduno/clmtrackr.

[9] http://www.menpo.org/menpofit/.

develop their own techniques or train existing ones with a larger number of points for identifying additional landmarks, which may escape the scope of the studies. Therefore, the lack of libraries able to identify such points in a more comprehensive and accurate manner and consider images under different conditions of acquisition is still a challenge to be overcome and can provide a unique opportunity for research.

An alternative to increasing the number of facial landmarks can be building models for facial components. Such models can improve the accuracy of the synthesis steps as in Xiong et al. (2010, 2007b), presented in Section 5.2. These models could be applied to the face as a whole and not only to the region of the eye.

Another alternative to landmark identification is a dense correspondence between two images, as shown in Yu et al. (2018). The dense correspondence estimation matches every pixel of one image with the equivalent pixel in another image, thus producing a greater amount of corresponding points between facial components.

10.1.3 Facial Alignment. The alignment of the faces allows the normalization of source and target images to be used as the basis for the synthesis. This normalization is important for images to have a better correspondence between facial components, which can improve the mapping of features from the source faces to the target face.

In most cases, alignment transformations were carried out on images (affine transformation, Procrustes transformation, and pixel-based correspondence), as presented in Table 5. The landmarks were used only as a basis for these transformations. However, in the case FAPU, normalization solely occurs in relation to landmarks.

A possible approach would be combining affine transformation with thin plate spline. This combination can enable more precise alignment of faces. A good alignment result is essential for using mapping approaches, especially using ERI.

10.2 Synthesis Methods

Table 6 illustrates the strengths and weaknesses of each of the types of image synthesis approaches analyzed. The studies were grouped according to the stages performed for the synthesis by their approaches. These stages represent each of the possible paths for Figure 3.

The studies analyzed in this SR aim at generating photorealistic images; however, not all studies will solve all the points related to this aspect. Table 3 shows that not all studies, for example, synthesize the inner region of the mouth and wrinkles. The synthesis on these regions eventually presents as secondary aspects in some studies, especially in earlier studies that focus on the synthesis of other regions of the face.

Most studies require different faces for target and source images. However, some approaches, such as blend shape (Section 7.1) and similar facial expression (Section 7.4) require images of the target face displaying different facial expressions. Such approaches are not good for all applications mentioned in Section 1, such as training/diagnosis of facial expressions, training data augmentation, and the like because they are more focused on other types of applications. Approaches that require multiple images of the target face are usually done by reenacting the facial expression from a source video, as shown in Thies et al. (2016).

Learning approaches (Section 6) have the advantage of defining those aspects that characterize the expression of an emotion because they synthesize expressive features based on more than one instance of the same facial expression. However, this requires a high processing load for building the model and also a large number of instances for more realistic results, especially for deep learning approaches (Zhou and Shi 2017). This can be problematic, especially in the case of special facial expressions in which there are few or just one example of the desired expression (Xiong

Table 6. Strengths and Weaknesses of the Various Synthesis Approaches Grouped According to the Stages Performed

Approach	Number of studies	Description	Strengths	Weaknesses
Training → Deformation	2% (1)	The face is deformed from the displacements of the facial landmarks learned.	By not treating directly the learning of pixels it decreases noise interference.	Largely dependent on the correctness of the location of landmarks.
Fusion → Training → Deformation → Mapping	2% (1)	The trained model is used to deform the image according to the moves learned. Finally, the lighting information is mapped.		
Deformation → Fusion → Training → Mapping	6% (3)	The face is deformed and decomposed. The training is carried out on the decomposed face. The faces generated by deformation and the trained models are merged. Finally, the lighting information is mapped.	Uses both the deformation and training results for the synthesis.	Suffers interference from deformation and mapping errors.
Fusion	25% (13)	The face is decomposed into identity and expression. An expression is reconstructed from the composition of the identity with the expression.	Separation of identity and expression for the reconstruction.	The reconstruction suffers interference from facial components that do not change, such as the hair.
Fusion → Deformation	2% (1)	The same above mentioned procedure (fusion) is applied, and the reconstructed face is then deformed to synthesize the expression.		
Fusion → Training	26% (14)	The training considers the separation of facial expression and identity subspaces for synthesis.	Considers a varied number of patterns as samples.	The facial expression can be synthesized incorrectly, especially when there are few examples of the facial expression desired.
Deformation → Mapping	21% (11)	The face is deformed to match the facial expression. The lighting information is then mapped to synthesize the facial expression.	The synthesis of the expressions occurs equally regardless of the number of examples of a particular facial expression.	The motion patterns can generate inconsistencies if the source and destination faces have facial components with considerable difference.
Deformation → Mapping → Muscles simulation	4% (2)	Uses the same approach to deformation and mapping, but muscle information is incorporated to improve the movement of landmarks or illumination mapping.	The found face already has a facial expression similar to the one desired.	Requires several images of the target face with different facial expressions.
Similar facial expression → Deformation → Mapping	4% (2)	Finds the target face with the facial expression most similar to the desired one, then performs the same approach of deformation and mapping.	Enhances synthesized images without relying on more examples.	Simulation can restrict real movements in a wrong way.
Clustering → Correlation → Deformation	2% (1)	Motion patterns are learned and correlated. This information is used to deform the face.	The same facial expression can have different movement patterns. The classification and correlation of these patterns allow synthesizing facial expression considering the related movement patterns.	The synthesized facial expression may not be desired one, if the resulting grouping is between two different facial expressions.
Blend shape	4% (2)	Images of the target face are blended by interpolation.	Synthesizes highly realistic images.	Requires several images of the target face.
Cloud model	2% (1)	The model can divide the facial expression features of the face for reconstruction with one source image.	Synthesizes highly realistic images.	Synthesized images may not have the desired degree of realism if the source face is not similar to the target face.

et al. 2010). The literature (Ekman and Rosenberg 1997) shows that some emotions are difficult to reproduce spontaneously and therefore are more difficult to represent in images. Often, when they are represented, the subject is asked to simulate the facial expression (Kanade et al. 2000), which can generate an artificial expression and, consequently, influence the desired realism of the target image. Furthermore, spontaneous facial expressions exhibit more subtle facial cues and are more easily classified into more than six basic expressions (Zhou and Shi 2017). Thus, they may require more examples for each facial expression.

On the other hand, approaches of deformation and mapping (Section 5) can reproduce any facial expression, even with just one example. However, the differences between the source and target faces can render the transfer of expressive characteristics impractical because there is no acceptable correspondence among the components of the two images. Consequently, the result can be extremely superficial. Some studies join deformation, mapping, and learning techniques to produce their approach, trying to make the best of each approach to synthesize the new image.

None of the studies discussed how the characteristics of the source images influenced the synthesized images. Lower-quality source images may result in a smaller degree of realism in synthesized images. Another factor that can influence the final result is the intensity of a particular facial expression since less expressive source faces tend to produce a less expressive target image. These problems mainly influence approaches of synthesis based on machine learning because these low-quality image features are incorporated into the model. Deformation and mapping approaches are less influenced because the selection of the source images is usually performed manually.

It is apparent that it is still a challenge to develop hybrid approaches that can aggregate the strengths of the classical approaches. Approaches should be developed to allow the synthesis of expressions through the adaptive use of source images, given the number and types of images available for reconstructing facial expressions.

In addition to the aspects shown in Table 6, there are other problems with these approaches, and these constitute research opportunities. An example is that fusion techniques do not consider the spatial relations of the data. The image is transformed into a one-dimensional vector for processing. Thus, the pixel relationships between neighbors are disregarded. Another open problem is the synthesis of facial expressions in uncontrolled environments, for example, with different lighting conditions and different poses. In the case of lighting, some papers try to create new models to reduce the influence of lighting variations (Xie et al. 2017; Xiong et al. 2010; Zhang et al. 2014). Such models are developed mainly because the mapping approach used (ERI) is sensitive to lighting variations.

The occlusion problem is only considered by Impett et al. (2014), which discusses the influence of artifacts on the images synthesized. Examples of common artifacts that influence the synthesis of facial expression are glasses and beards. None of the studies analyzed deals with removing the influence of occlusion in the final result. Occlusion removal (or reconstruction) approaches could also be applied for synthesizing facial expressions.

The facial expressiveness of every kind of emotion is not universal. That is, different people express the same emotion in different ways (see Section 1). Even similar neutral faces have different facial expressions (Zhang and Wei 2012). Therefore, an open challenge is the development of approaches capable of efficiently considering this variability.

Finally, another aspect little considered by the reviewed articles was the synthesis of facial expressions in different poses. Only Xie et al. (2018) and Zhou and Shi (2017) show examples of synthesized facial expressions with the source and target images in different poses. This may be because most approaches are sensitive to variations in pose in all the stages of the process: training, deformation, and especially in illumination mapping (ERI). Although the work of Xie et al. (2018) and Zhou and Shi (2017) brings great advances to illumination mapping even in different poses, it requires the manual extraction of the regions of wrinkles. A possible alternative would be the use of approaches to normalize poses on the faces, as shown in Cole et al. (2017).

In this review, we analyzed approaches for facial expression synthesis considering 2D techniques for image synthesis. Some works use 3D techniques for the synthesis of facial expressions in photographs. These papers were not included because they are outside the scope of this SR. However, they address important approaches that deserve to be cited in order to improve reflections about challenges and opportunities in this research area. For instance, Song et al. (2007) proposes a geometry encoding approach for representing the face in a 3D surface, thus enabling facial expression synthesis in both 2D and 3D cases. Another example is Face2Face from Thies et al. (2016) that reconstructs face identity using a model based on a multilinear PCA and then deforms the mesh vertices by blending deformations. There are other important approaches that did not directly address the facial expression synthesis. This is the case of Zhang et al. (2017), which proposes an age progression/regression approach that considers different facial expressions.

10.3 Metrics for the Evaluation of the Results

The aspects examined by the authors in the synthesized images (Section 8.3) highlight important issues for evaluating results. The use of an evaluation metric can contribute to reducing the analysis bias of synthesized facial expressions. Of the studies analyzed (Table 3), only a small part used an objective (Section 8.1) or subjective (Section 8.2) metric to evaluate the results.

Of the two types of metrics analyzed, objective assessments were the most widely used. The objective evaluation is carried out systematically and quantitatively, applying metrics that do not depend on the perception of a person. It is thus possible to make the results replicable and easier to compare with other studies. Therefore, these metrics, in addition to evaluating results, can serve as a basis for comparison between studies. However, no metrics used in objective assessment can actually be considered a standard. Indeed, the few studies that presented objective evaluation used different metrics. Thus, deep studies on appropriate metrics and even proposing new metrics could be a research contribution to the area.

Subjective evaluation assesses the results by checking a user's perception of the synthesized images. This perception is important to ensure that the synthesized images are convincing to the eyes of those who use them. Despite being relatively simple, this approach poses the challenge of requiring a considerable number of users to evaluate the results in addition to the need to build appropriate experiments without tiring the user (Wang and Bovik 2006). The systematic description of experiment protocols was not verified in the articles analyzed. Thus, the study of protocols for conducting experiments could contribute to improving the state of the art of this area of research.

Creating metrics and protocols for assessing results requires standardization and the availability of some aspects. It is necessary to define which set of evaluation metrics is important to evaluate each aspect. There should also be a number of synthesized images available for comparative purposes. As in other areas, such as Software Engineering (Andrews et al. 2005; Sjoeberg et al. 2005; Thelin et al. 2003), the creation of repositories to provide programs and benchmarks for evaluating new methods is a challenge, but it can become an important contribution to the research of facial images synthesis. For a more complete evaluation, objective and subjective metrics could be used complementarily, thus creating an approach that considers both the user's perception and systematic quantification of the results.

The problem that objective comparative approaches do not consider the relationship between image regions (Table 7) can be circumvented by checking other elements rather than just the vector of pixels. For example, the consistency of the synthesized shape compared to the landmarks of ground truth could be checked. The approach for verifying the consistency of landmarks

Table 7. Strengths and Weaknesses in the Metrics for the Evaluation of the Results

Type of evaluation	Metric	Number of studies	Strengths	Weaknesses
Objective	Comparison with the ground truth image	9% (5)	Checks if the image is closest to the actual situation.	The pixels of the images are compared directly. Does not consider the relationship between them. Can also cause inconsistencies in the case of noise in the compared image.
	Comparison with the source image	8% (4)	Checks if the desired expression was mimic well	
	Comparison with the target neutral image	4% (2)	Checks if a face identity has not been lost.	
	Comparison with images from other approaches	9% (5)	Checks the improvements to the state of the art.	
	Consistency of landmarks	2% (1)	Checks the consistency of the shape without using comparative images.	The acceptable range may allow inconsistent facial expressions.
	Facial expression recognition	4% (2)	Checks the facial expression automatic and without using comparative images.	The facial expression recognition algorithm may have a error itself.
	No-reference	2% (2)	Checks if any noises and blocks have been introduced without using comparative images.	Does not consider the specificities of the face and facial expression.
Subjective	Images realism	11% (6)	Checks the quality and consistency of expressive images.	Does not indicate what is wrong with the image.
	Facial expression identification	9% (5)	Checks if the synthesized facial expression is consistent with the proposed expression.	A certain facial expression can be understood differently according to each evaluator.
	Pearson identity preservation	4% (2)	Checks if a face identity has not been lost.	The task of recognition can be difficult to identify people of different ethnicities.

(Section 8.1) checks whether the landmarks of synthesized images are within a reasonable range. That is, it does not use the landmarks of ground truth for validation; the comparison with the ground truth could better verify the consistency of the shape.

Some improvements could also be made in subjective assessment tasks. For example, to evaluate the realism of the images (Section 8.2), the evaluator could list the inconsistencies found in synthesized images. In the case of the facial expression identification task (Section 8.2), among a set of expressive images, the one that best resembled the synthesized image could be chosen. Thus, the problem of the synthesized images evaluator not knowing the name of a certain facial expression could be avoided.

The automatic evaluation of face recognition is a good alternative to a subjective evaluation of facial expression since the algorithms used have high accuracy (Zhou and Shi 2017) and can even surpass human-level evaluation (Bartlett et al. 2014; Janssen et al. 2014). However, people more easily recognize facial expressions tied to the context in which the emotion is inserted (Crivelli and Fridlund 2018; Ekman and O'Sullivan 1988; Fridlund 1991), which can make the comparison unfair. While objective assessments allow comparing metrics, the subjective ones provide the perception of a final user regarding the emotion.

None of the subjective assessments of the facial expressions identification tasks considers the facial expression within a context. Generally, evaluators should classify the facial expression on one of the provided labels. An alternative is to provide a story instead of a label. Another option is to provide real images of a person performing each expression and ask the evaluator to relate the synthesized image with one of the emotions, which can reduce the bias associated with the label.

10.4 Facial Expression Image Database

Source images are important for most of the approaches used for synthesizing facial expressions. Facial expression databases contain images based on some requirements for their possible applications, including synthesis. Thus, these databases capture the facial expression images necessary to meet a particular characteristic or to create a new expression database (e.g., for a study of new approaches to profiles and facial expression).

One of the challenges in synthesis is generating the different types of facial expressions (emotions, blinking, talking, etc.) possible for a face. This task becomes more difficult when it is necessary to synthesize a different emotion from the six universal ones (Section 1). Some facial expressions only have a few examples; therefore, they are more challenging. An attempt to cover the largest possible number of facial expressions is the CK+, which performs this task by providing examples of various units of action of the FACS.

Facial expression databases do not generally provide information on how expressive a facial expression of an emotion is. The availability of such information could be very useful for synthesizing the same facial expression at different intensities.

Of the databases presented in the articles analyzed, the most widely used were CK/CK+ and JAFFE. This is probably a consequence of these databases being more complete (addressing the six universal emotions), classifying the emotions, being free of costs, and of longer standing. Between the two, the most widely used was CK/CK+. This may be due to the breadth of different ethnicities and both genders being available. Yet the CK/CK+ database contains two serious problems: (i) the images are mostly gray, which can be a problem when generating a new facial expression for colored images, and (ii) the resolution of the images is 640 × 490, which can lead to less photorealistic results.

Table 4 shows that a large portion of the requirements for synthesizing facial expressions are fulfilled. However, the factors listed in the previous paragraphs show possibilities for improving these databases. It should be kept in mind that only the databases used by the articles included in this review were analyzed. Other databases may be considered better for facial expression synthesis purposes. For example, the images from the databases proposed in Aifanti et al. (2010), Mavadati et al. (2013), Pantic et al. (2005), and Sneddon et al. (2012) are colored, have higher resolution, and use facial expression labels. Thus, an in-depth study on the theme could consider a larger number of databases and a complete analysis of the topic for facial expression synthesis.

11 CONCLUSION

We conducted a systematic literature search to analyze existing approaches to the synthesis of facial expressions in photographs, as well as additional elements that influence the results of these approaches: preprocessing techniques, facial expressions databases, and the different forms and metrics for evaluating results.

The synthesis of facial expressions is a challenging topic. A precise approach should allow generating a new facial expression consistently and realistically while preserving the identity of the face. Despite being an area relatively well explored in the literature, some challenges still remain, such as handling occlusions, pose variations, and changes in lighting conditions and a lack of standards and protocols for evaluating the results.

The challenges mentioned are important research themes to be exploited, which can generate significant contributions to this area. Among these themes are the development of hybrid approaches, the creation of repositories with benchmarks, and the establishment of protocols and standardized metrics for evaluation. The synthesis of facial expressions in photographs has several applications. Contining research in this area can have social impact as it could have suitable applications in the therapy of patients with psychiatric disorders, biometric systems that are invariant to facial expressions, and facial animation, among others.

ACKNOWLEDGMENTS

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001 – and the Brazilian Institute of Science and Technology in Medicine Assisted by Scientific Computing (INCT-MACC).

REFERENCES

B. Abboud and F. Davoine. 2004. Appearance factorization based facial expression recognition and synthesis. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04).* Vol. 4. 163–166. DOI: https://doi.org/10.1109/ICPR.2004.1333729

B. Abboud and F. Davoine. 2005. Bilinear factorisation for facial expression analysis and synthesis. *IEE Proceedings on Vision, Image and Signal Processing* 152, 3 (June 2005), 327–333. DOI: https://doi.org/10.1049/ip-vis:20045060

Bouchra Abboud, Franck Davoine, and Mô Dang. 2004. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication* 19, 8 (2004), 723–740. DOI: https://doi.org/10.1016/j.image.2004.05.009

Swapna Agarwal, Moitreya Chatterjee, and Dipti Prasad mukherjee. 2012. Synthesis of emotional expressions specific to facial structure. In *Proceedings of the 8th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'12).* ACM, New York, Article 28, 8 pages. DOI: https://doi.org/10.1145/2425333.2425361

N. Aifanti, C. Papachristou, and A. Delopoulos. 2010. The MUG facial expression database. In *Proceedings of the11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'10).* 1–4.

Joan Alabort-i Medina, Epameinondas Antonakos, James Booth, Patrick Snape, and Stefanos Zafeiriou. 2014. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM'14).* ACM, New York, 679–682. DOI: https://doi.org/10.1145/2647868.2654890

B. Amberg, R. Knothe, and T. Vetter. 2008. Expression invariant 3D face recognition with a morphable model. In *Proceedings of the 8th IEEE International Conference on Automatic Face Gesture Recognition.* 1–6. DOI: https://doi.org/10.1109/AFGR.2008.4813376

J. H. Andrews, L. C. Briand, and Y. Labiche. 2005. Is mutation an appropriate tool for testing experiments?. In *Proceedings of the 27th International Conference on Software Engineering (ICSE'05).* ACM, New York, 402–411. DOI: https://doi.org/10.1145/1062455.1062530

Aristotle, William David Ross, John Alexander Smith, T. Loveday, L. D. Dowdall, Edward Seymour Forster, and Harold Henry Joachim. 1913. *The Works of Aristotle: Opuscula.* Oxford University Press.

T. Baltrusaitis, P. Robinson, and L. P. Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops.* 354–361. DOI: https://doi.org/10.1109/ICCVW.2013.54

Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, and Kang Lee. 2014. Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology* 24, 7 (31 Mar 2014), 738–743. DOI: https://doi.org/10.1016/j.cub.2014.02.009

Vinay Bettadapura. 2012. Face expression recognition and analysis: The state of the art. *CoRR* abs/1203.6722 (2012). arxiv:1203.6722 http://arxiv.org/abs/1203.6722

Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. 2003. Reanimating faces in images and video. *Computer Graphics Forum*, Vol. 22. Wiley Online Library, 641–650.

Steven M. Boker, Jeffrey F. Cohn, Barry-John Theobald, Iain Matthews, Timothy R. Brick, and Jeffrey R. Spies. 2009. Effects of damping head movement and facial expression in dyadic conversation using real–time facial expression tracking and synthesized avatars. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364, 1535 (2009), 3485–3495.

X. Chao and F. Zhiyong. 2008. Facial expression recognition and synthesis on affective emotions composition. In *Proceedings of the International Seminar on Future BioMedical Information Engineering.* 144–147. DOI: https://doi.org/10.1109/FBIE.2008.53

Yufang Cheng and Shuhui Ling. 2008. 3D animated facial expression and autism in Taiwan. In *Proceedings of the 8th IEEE International Conferenceon Advanced Learning Technologies (ICALT'08).* IEEE, 17–19.

Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2017. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR* abs/1711.09020 (2017). arxiv:1711.09020 http://arxiv.org/abs/1711.09020

F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. 2017. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3386–3395. DOI: https://doi.org/10.1109/CVPR.2017.361

Carlos Crivelli and Alan J. Fridlund. 2018. Facial displays are tools for social influence. *Trends in Cognitive Sciences* 22, 5 (2018), 388–399. DOI: https://doi.org/10.1016/j.tics.2018.02.006

Charles Darwin. 1916. *The Expression of the Emotions in Man and Animals*. D. Appleton and Co., 406 pages.

Yangzhou Du and Xueyin Lin. 2002. Mapping emotional status to facial expressions. In *Proceedings of the 16th International Conference on Pattern Recognitionon*, Vol. 2. 524–527. DOI: https://doi.org/10.1109/ICPR.2002.1048355

Yangzhou Du and Xueyin Lin. 2003. Emotional facial expression model building. *Pattern Recognition Letters* 24, 16 (2003), 2923–2934. DOI: https://doi.org/10.1016/S0167-8655(03)00153-3

Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124.

Paul Ekman and Wallace V. Friesen. 1975. *Pictures of Facial Affect*. consulting psychologists press.

Paul Ekman and Wallace V. Friesen. 1976. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior* 1, 1 (1976), 56–75.

Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth. 1972. *Emotion in the Human Face: Guidelines for Research and An integration of Findings*. Pergamon Press, Oxford, England. 191 pages.

Paul Ekman and Maureen O'Sullivan. 1988. The role of context in interpreting facial expression: Comment on Russell and Fehr (1987). *Journal of Experimental Psychology: General* 117, 1 (1988), 86–88. DOI: https://doi.org/10.1037/0096-3445.117.1.86

Paul Ekman and Erika L. Rosenberg. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Nikolaos Ersotelos and Feng Dong. 2008. Building highly realistic facial modeling and animation: A survey. *The Visual Computer* 24, 1 (2008), 13–30. DOI: https://doi.org/10.1007/s00371-007-0175-y

Alan J. Fridlund. 1991. Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology* 60, 2 (1991), 229–240. DOI: https://doi.org/10.1037/0022-3514.60.2.229

Yun Fu, N. N. Zheng, Ting Zhang, and F. Zhuo. 2003. Facial expression transformation, aging and invisible view reconstruction. *Acta Electronica Sinica* 31, 12A (2003), 1955–1970.

Hiroki Fujishiro, Takanori Suzuki, Shinya Nakano, Akinobu Mejima, and Shigeo Morishima. 2009. A natural smile synthesis from an artificial smile. In *SIGGRAPH'09: Posters*. ACM, New York, Article 59, 1 pages. DOI: https://doi.org/10.1145/1599301.1599360

John Ghent and John McDonald. 2005. Photo-realistic facial expression synthesis. *Image and Vision Computing* 23, 12 (2005), 1041–1050. DOI: https://doi.org/10.1016/j.imavis.2005.06.011

Ofer Golan and Simon Baron-Cohen. 2006. Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia. *Development and Psychopathology* 18 (2006), 591–617. DOI: https://doi.org/10.1017/S0954579406060305

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing*, Minho Lee, Akira Hirose, Zeng-Guang Hou, and Rhee Man Kil (Eds.). Springer Berlin Heidelberg, 117–124.

Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-PIE. *Image and Vision Computing* 28, 5 (2010), 807–813. DOI: https://doi.org/10.1016/j.imavis.2009.08.002 Best of Automatic Face and Gesture Recognition 2008.

Ouriel Grynszpan, Jean-Claude Martin, and Jacqueline Nadel. 2008. Multimedia interfaces for users with high functioning autism: An empirical investigation. *International Journal of Human-Computer Studies* 66, 8 (2008), 628–639.

Madeline B, Harms, Alex Martin, and Gregory L. Wallace. 2010. Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychology Review* 20, 3 (2010), 290–322.

D. He, J. Tang, and X. Jiing. 2008. Facial expression synthesis based on expressional feature space and expression decomposition. In *Proceedings of the International Conference on Computer Science and Software Engineering*, Vol. 1. 1009–1012. DOI: https://doi.org/10.1109/CSSE.2008.1010

Ursula Hess. 2001. The communication of emotion. *Emotions, Qualia and Consciousness* (2001), 397–409.

Leonardo Impett, Peter Robinson, and Tadas Baltrusaitis. 2014. A facial affect mapping engine. In *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces (IUI Companion'14)*. ACM, New York, 33–36. DOI: https://doi.org/10.1145/2559184.2559203

Carroll E. Izard. 1971. *The Face of Emotion.*Appleton-Century-Crofts. 468 pages.

Rachael E. Jack, Wei Sun, Ioannis Delis, Oliver G. B. Garrod, and Philippe G. Schyns. 2016. Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General* 145, 6 (2016), 708–730. DOI: https://doi.org/10.1037/xge0000162

Joris H. Janssen, Wijnand A. IJsselsteijn, and Joyce H. D. M. Westerink. 2014. How affective technologies can influence intimate interactions and improve social connectedness. *International Journal of Human-Computer Studies* 72, 1 (2014), 33–43. DOI: https://doi.org/10.1016/j.ijhcs.2013.09.007

J. Jia, S. Zhang, and L. Cai. 2010. Facial expression synthesis based on motion patterns learned from face database. In *Proceedings of the 2010 IEEE International Conference on Image Processing*. 3973–3976. DOI: https://doi.org/10.1109/ICIP.2010.5653914

T. Kanade, J. F. Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings of teh 4th IEEE International Conference on Automatic Face and Gesture Recognition*. 46–53. DOI: https://doi.org/10.1109/AFGR.2000.840611

Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. IEEE Computer Society, Washington, DC, 1867–1874. DOI: https://doi.org/10.1109/CVPR.2014.241

Erwin Keeve, Sabine Girod, Ron Kikinis, and Bernd Girod. 1998. Deformable modeling of facial tissue for craniofacial surgery simulation. *Computer Aided Surgery* 3, 5 (1998), 228–238.

A. Z. Kouzani. 1999. Facial expression synthesis. In *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, Vol. 1. 643–647. DOI: https://doi.org/10.1109/ICIP.1999.821713

Uttama Lahiri, Esubalew Bekele, Elizabeth Dohrmann, Zachary Warren, and Nilanjan Sarkar. 2013. Design of a virtual reality based adaptive response technology for children with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 21, 1 (2013), 55–64.

Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. 2010. Presentation and validation of the radboud faces database. *Cognition and Emotion* 24, 8 (2010), 1377–1388. DOI: https://doi.org/10.1080/02699930903485076

Chan-Su Lee and A. Elgammal. 2006. Nonlinear shape and appearance models for facial expression analysis and synthesis. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 1. 497–502. DOI: https://doi.org/10.1109/ICPR.2006.867

J. H. Lee, J. M. Lee, H. J. Kim, and Y. S. Moon. 2007. Automatic synthesis of realistic facial expressions. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*. 46–51. DOI: https://doi.org/10.1109/ISSPIT.2007.4458169

J. H. Lee, Ki Tae Park, and Y. S. Moon. 2009. Realistic expression mapping robust to various lighting conditions. In *2009 Digest of Technical Papers International Conference on Consumer Electronics*. 1–2. DOI: https://doi.org/10.1109/ICCE.2009.5012333

M. Y. Y. Leung, Hung Yen Hui, and I. King. 1996. Facial expression synthesis by radial basis function network and image warping. In *Proceedings of the IEEE International Conferenceon Neural Networks*, Vol. 3. 1400–1405. DOI: https://doi.org/10.1109/ICNN.1996.549104

K. Li, Q. Dai, R. Wang, Y. Liu, F. Xu, and J. Wang. 2014. A data-driven approach for facial expression retargeting in video. *IEEE Transactions on Multimedia* 16, 2 (Feb 2014), 299–310. DOI: https://doi.org/10.1109/TMM.2013.2293064

K. Li, F. Xu, J. Wang, Q. Dai, and Y. Liu. 2012. A data-driven approach for facial expression synthesis in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 57–64. DOI: https://doi.org/10.1109/CVPR.2012.6247658

Shan Li and Weihong Deng. 2018. Deep facial expression recognition: A survey. *CoRR* abs/1804.08348 (2018). arxiv:1804.08348 http://arxiv.org/abs/1804.08348

Xin Li, Chieh-Chih Chang, and Shi-Kuo Chang. 2007. Face alive icon. *Journal of Visual Languages & Computing* 18, 4 (2007), 440–453. DOI: https://doi.org/10.1016/j.jvlc.2007.02.008 Visual Interactions in Software Artifacts.

H. Liang, R. Liang, M. Song, and X. He. 2016. Coupled dictionary learning for the detail-enhanced synthesis of 3-D facial expressions. *IEEE Transactions on Cybernetics* 46, 4 (April 2016), 890–901. DOI: https://doi.org/10.1109/TCYB.2015.2417211

Zicheng Liu, Ying Shan, and Zhengyou Zhang. 2001. Expressive expression mapping with ratio images. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*. ACM, New York, 271–276. DOI: https://doi.org/10.1145/383259.383289

P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. 2010. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 94–101. DOI: https://doi.org/10.1109/CVPRW.2010.5543262

M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. 1998. Coding facial expressions with gabor wavelets. In *Proceedings of the 3rd International Conference on Face & Gesture Recognition (FG'98)*. IEEE Computer Society, Washington, DC, 200–205.

I. Macedo, E. V. Brazil, and L. Velho. 2006. Expression transfer between photographs through multilinear AAM's. In *Proceedings of the 19th Brazilian Symposium on Computer Graphics and Image Processing*. 239–246. DOI: https://doi.org/10.1109/SIBGRAPI.2006.18

S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. 2013. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing* 4, 2 (April 2013), 151–160. DOI: https://doi.org/10.1109/T-AFFC.2013.4

Dhwani Mehta, Mohammad Faridul Haque Siddiqui, and Ahmad Y. Javaid. 2018. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors* 18, 2 (2018). DOI: https://doi.org/10.3390/s18020416

E. Mendi and C. Bayrak. 2011. Facial animation framework for web and mobile platforms. In *Proceedings of the IEEE 13th International Conference on e-Health Networking, Applications and Services*. 52–55. DOI: https://doi.org/10.1109/HEALTH.2011.6026785

K. Messer, J. Matas, J. Kittler, and K. Jonsson. 1999. XM2VTSDB: The extended M2VTS database. In *Proceedings of the 2nd International Conference on Audio and Video-based Biometric Person Authentication*. 72–77.

Daisuke Mima, Hiroyuki Kubo, Akinobu Maejima, and Shigeo Morishima. 2011. Automatic generation of facial wrinkles according to expression changes. In *SIGGRAPH Asia 2011 Posters (SA'11)*. ACM, New York, Article 1, 1 pages. DOI: https://doi.org/10.1145/2073304.2073306

Saeed Montazeri Moghadam and Seyyed Ali Seyyedsalehi. 2018. Nonlinear analysis and synthesis of video images using deep dynamic bottleneck neural networks for face recognition. *Neural Networks* (2018). DOI: https://doi.org/10.1016/j.neunet.2018.05.016

A. Mohammadian, H. Aghaeinia, and F. Towhidkhah. 2016. Diverse videos synthesis using manifold-based parametric motion model for facial understanding. *IET Image Processing* 10, 4 (2016), 253–260. DOI: https://doi.org/10.1049/iet-ipr.2014.0905

Umar Mohammed, Simon J. D. Prince, and Jan Kautz. 2009. Visio-lization: Generating novel facial images. *ACM Transactions on Graphics* 28, 3, Article 57 (July 2009), 8 pages. DOI: https://doi.org/10.1145/1531326.1531363

Maryam Moosaei, Cory J. Hayes, and Laurel D. Riek. 2015. Facial expression synthesis on robots: An ROS module. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts (HRI'15 Extended Abstracts)*. ACM, New York, 169–170. DOI: https://doi.org/10.1145/2701973.2702053

Jun-yong Noh and Ulrich Neumann. 2001. Expression cloning. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*. ACM, New York, 277–288. DOI: https://doi.org/10.1145/383259.383290

M. M. Nordstrøm, M. Larsen, J. Sierakowski, and M. B. Stegmann. 2004. *The IMM Face Database - An Annotated Dataset of 240 Face Images*. Technical Report. Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby. http://www2.imm.dtu.dk/pubdb/p.php?3160.

M. Pantic, M. Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*. DOI: https://doi.org/10.1109/ICME.2005.1521424

Mark Petticrew and Helen Roberts. 2008. *Systematic Reviews in the Social Sciences: A Practical Guide*. John Wiley & Sons.

Karen Pierce, Frank Haist, Farshad Sedaghat, and Eric Courchesne. 2004. The brain response to personally familiar faces in autism: Findings of fusiform activity and beyond. *Brain* 127, 12 (2004), 2703–2716. DOI: https://doi.org/10.1093/brain/awh289

Cristiana Castanho de Almeida Rocca, Eveline van den Heuvel, Sheila C Caetano, and Beny Lafer. 2009. Facial emotion recognition in bipolar disorder: A critical review. *Revista Brasileira de Psiquiatria* 31, 2 (2009), 171–180.

James A. Russell. 1994. Is there universal recognition of emotion from facial expression?: A review of the cross-cultural studies. *Psychological Bulletin* 115, 1 (1994), 102–141. DOI: https://doi.org/10.1037/0033-2909.115.1.102

M. Sabzevari, S. Toosizadeh, S. R. Quchani, and V. Abrishami. 2010. A fast and accurate facial expression synthesis system for color face images using face graph and deep belief network. In *Proceedings of the International Conference on Electronics and Information Engineering (ICEIE)*, Vol. 2. 354–358. DOI: https://doi.org/10.1109/ICEIE.2010.5559797

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2011b. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision* 91, 2 (Jan. 2011), 200–215. DOI: https://doi.org/10.1007/s11263-010-0380-4

J. M. Saragih, S. Lucey, and J. F. Cohn. 2011a. Real-time avatar animation from a single image. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG'11)*. 117–124. DOI: https://doi.org/10.1109/FG.2011.5771383

M. Seo and Y. W. Chen. 2012. Two-step subspace learning for texture synthesis of facial images. In *Proceedings of the 6th International Conference on Information Science and Service Science and Data Mining (ISSDM)*. 483–486.

Zhixin Shu, Lei Huang, and Changping Liu. 2013. 3d facial expression synthesis from a single image using a model set. In *Proceedings of the Asian Conference on Computer Vision*. Springer, 272–283.

T. Sim, S. Baker, and M. Bsat. 2003. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 12 (Dec 2003), 1615–1618. DOI: https://doi.org/10.1109/TPAMI.2003.1251154

D. I. K. Sjoeberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N. K. Liborg, and A. C. Rekdal. 2005. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering* 31, 9 (Sept 2005), 733–753. DOI: https://doi.org/10.1109/TSE.2005.97

I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. 2012. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing* 3, 1 (Jan 2012), 32–41. DOI: https://doi.org/10.1109/T-AFFC.2011.26

M. Song, Z. Dong, C. Theobalt, H. Wang, Z. Liu, and H. P. Seidel. 2007. A generic framework for efficient 2-D and 3-D facial expression analogy. *IEEE Transactions on Multimedia* 9, 7 (Nov 2007), 1384–1395. DOI: https://doi.org/10.1109/TMM.2007.906591

M. Song, D. Tao, Z. Liu, X. Li, and M. Zhou. 2010. Image ratio features for facial expression recognition application. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40, 3 (June 2010), 779–788. DOI: https://doi.org/10.1109/TSMCB.2009.2029076

M. Song, H. Wang, J. Bu, C. Chen, and Z. Liu. 2006. Subtle facial expression modeling with vector field decomposition. In *Proceedings of the International Conference on Image Processing*. 2101–2104. DOI: https://doi.org/10.1109/ICIP.2006.312822

T. F. Su, J. J. Li, C. H. Duan, S. F. Wang, and S. H. Lai. 2011. Parallelized face based RMS system on a multi-core embedded computing platform. In *Proceedings of the 40th International Conference on Parallel Processing Workshops*. 199–206. DOI: https://doi.org/10.1109/ICPPW.2011.52

H. C. Tan, Hao Chen, Wu-Hong Wang, and Jian-Wei Shi. 2009. Incremental tensor by face synthesis estimating for face recognition. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, Vol. 6. 3129–3133. DOI: https://doi.org/10.1109/ICMLC.2009.5212704

C. Tay, M. Obaid, R. Mukundan, and A. Bainbridge-Smith. 2009. Facial expressions using a quadratic deformation model: Analysis and synthesis. In *Proceedings of the 24th International Conference Image and Vision Computing New Zealand*. 50–55. DOI: https://doi.org/10.1109/IVCNZ.2009.5378363

Stephan F. Taylor and Angus W. MacDonald. 2012. Brain mapping biomarkers of socio-emotional processing in schizophrenia. *Schizophrenia Bulletin* 38, 1 (2012), 73–80.

R. L. Testa, A. H. N. Muniz, L. U. S. Carpio, R. d. S. Dias, C. C. d. A. Rocca, A. M. Lima, and F. d. L. d. S. N. Marques. 2015. Generating facial emotions for diagnosis and training. In *Proceedings of the IEEE 28th International Symposium on Computer-Based Medical Systems*. 304–309. DOI: https://doi.org/10.1109/CBMS.2015.59

T. Thelin, P. Runeson, and C. Wohlin. 2003. An experimental comparison of usage-based and checklist-based reading. *IEEE Transactions on Software Engineering* 29, 8 (Aug 2003), 687–704. DOI: https://doi.org/10.1109/TSE.2003.1223644

J. Thies, M. ZollhÃűfer, M. Stamminger, C. Theobalt, and M. Nieasner. 2016. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395. DOI: https://doi.org/10.1109/CVPR.2016.262

Frank Wallhoff. 2004. Fgnet-facial expression and emotion database. *Technische Universität München* (2004).

Frank Wallhoff. 2006. Database with Facial Expressions and Emotions from Technical University of Munich (FEEDTUM). http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html.

Guoyin Wang, Changlin Xu, Qinghua Zhang, and Xiaorong Wang. 2012. *A Multi-step Backward Cloud Generator Algorithm*. Springer Berlin Heidelberg, 313–322. DOI: https://doi.org/10.1007/978-3-642-32115-3_37

Hongcheng Wang and N. Ahuja. 2003. Facial expression decomposition. In *Proceedings of the 9th IEEE International Conference on Computer Vision*. 958–965 vol.2. DOI: https://doi.org/10.1109/ICCV.2003.1238452

Hao Wang and Kongqiao Wang. 2008. Affective interaction based on person-independent facial expression space. *Neurocomputing* 71, 10–12 (2008), 1889–1901. DOI: https://doi.org/10.1016/j.neucom.2007.10.022 Neurocomputing for Vision ResearchAdvances in Blind Signal Processing.

Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. 2017. Facial feature point detection: A comprehensive survey. *Neurocomputing* (2017). DOI: https://doi.org/10.1016/j.neucom.2017.05.013

Zhou Wang and Alan C. Bovik. 2006. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing* 2, 1 (2006), 1–156.

Zhou Wang, H. R. Sheikh, and A. C. Bovik. 2002. No-reference perceptual quality assessment of JPEG compressed images. In *Proceedings of the International Conference on Image Processing*, Vol. 1. I–477–I–480. DOI: https://doi.org/10.1109/ICIP.2002.1038064

Wei Wei, Chunna Tian, Stephen John Maybank, and Yanning Zhang. 2016. Facial expression transfer method based on frequency analysis. *Pattern Recognition* 49 (2016), 115–128. DOI: https://doi.org/10.1016/j.patcog.2015.08.004

J. Wu, H. Chi, S. Wang, and L. Chi. 2010. Facial expression synthesis based on cloud model. In *Proceedings of the 2nd International Workshop on Intelligent Systems and Applications*. 1–4. DOI: https://doi.org/10.1109/IWISA.2010.5473397

T. Wu, N. J. Butko, P. Ruvulo, M. S. Bartlett, and J. R. Movellan. 2009. Learning to make facial expressions. In *Proceedings of the IEEE 8th International Conference on Development and Learning*. 1–6. DOI: https://doi.org/10.1109/DEVLRN.2009.5175536

W. Xie, L. Shen, and J. Jiang. 2017. A novel transient wrinkle detection algorithm and its application for expression synthesis. *IEEE Transactions on Multimedia* 19, 2 (Feb 2017), 279–292. DOI:https://doi.org/10.1109/TMM.2016.2614429

W. Xie, L. Shen, M. Yang, and Q. Hou. 2015. Lighting difference based wrinkle mapping for expression synthesis. In *Proceedings of the 8th International Congress on Image and Signal Processing (CISP)*. 636–641. DOI:https://doi.org/10.1109/CISP.2015.7407956

Weicheng Xie, Linlin Shen, Meng Yang, and Jianmin Jiang. 2018. Facial expression synthesis with direction field preservation based mesh deformation and lighting fitting based wrinkle mapping. *Multimedia Tools and Applications* 77, 6 (1 Mar 2018), 7565–7593. DOI:https://doi.org/10.1007/s11042-017-4661-6

Lei Xiong, Nanning Zheng, Shaoyi Du, and Lan Wu. 2009. Extended facial expression synthesis using statistical appearance model. In *Proceedings of the 4th IEEE Conference on Industrial Electronics and Applications*. 1582–1587. DOI:https://doi.org/10.1109/ICIEA.2009.5138461

Lei Xiong, Nanning Zheng, Jianyi Liu, Shaoyi Du, and Yuehu Liu. 2010. Eye synthesis using the eye curve model. *Image and Vision Computing* 28, 3 (2010), 329–342. DOI:https://doi.org/10.1016/j.imavis.2009.06.001

L. Xiong, N. Zheng, Q. You, and J. Liu. 2007a. Facial expression sequence synthesis based on shape and texture fusion model. In *Proceedings of theIEEE International Conference on Image Processing*, Vol. 4. 473–476. DOI:https://doi.org/10.1109/ICIP.2007.4380057

L. Xiong, N. Zheng, Q. You, J. Liu, and S. Du. 2007b. Eye synthesis using the eye curve model. In *Proceedings of the19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Vol. 2. 531–534. DOI:https://doi.org/10.1109/ICTAI.2007.84

Yang Yang, Nanning Zheng, Yuehu Liu, Lei Yang, Ping Wei, and Y. Nishio. 2009. Interactive facial sketch expression generation using local constraints. In *Proceedings of the IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS'09)*, Vol. 2. 864–868. DOI:https://doi.org/10.1109/ICICISYS.2009.5358275

Wei Yu, Xiaoshuai Sun, Kuiyuan Yang, Yong Rui, and Hongxun Yao. 2018. Hierarchical semantic image matching using CNN feature pyramid. *Computer Vision and Image Understanding* 169 (2018), 40–51. DOI:https://doi.org/10.1016/j.cviu.2018.01.001

Qingshan Zhang, Zicheng Liu, Baining Guo, and Harry Shum. 2003. Geometry-driven photorealistic facial expression synthesis. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA'03)*. Eurographics Association, Aire-la-Ville, Switzerland, 177–186.

Qingshan Zhang, Z. Liu, Gaining Quo, D. Terzopoulos, and Heung-Yeung Shum. 2006. Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Computer Graphics* 12, 1 (Jan 2006), 48–60. DOI:https://doi.org/10.1109/TVCG.2006.9

Shen Zhang, Zhiyong Wu, Helen M Meng, and Lianhong Cai. 2010. Facial expression synthesis based on emotion dimensions for affective talking avatar. In *Modeling Machine Emotions for Realizing Intelligence*. Springer, 109–132.

Y. Zhang, W. Lin, B. Sheng, J. Wu, H. Li, and C. Zhang. 2012. Facial expression mapping based on elastic and muscle-distribution-based models. In *Proceedings of the IEEE International Symposium on Circuits and Systems*. 2685–2688. DOI:https://doi.org/10.1109/ISCAS.2012.6271860

Yihao Zhang, Weiyao Lin, Bing Zhou, Zhenzhong Chen, Bin Sheng, and Jianxin Wu. 2014. Facial expression cloning with elastic and muscle models. *Journal of Visual Communication and Image Representation* 25, 5 (2014), 916–927. DOI:https://doi.org/10.1016/j.jvcir.2014.02.010

Yanning Zhang and Wei Wei. 2012. A realistic dynamic facial expression transfer method. *Neurocomputing* 89 (2012), 21–29. DOI:https://doi.org/10.1016/j.neucom.2012.01.019

Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. *CoRR* abs/1702.08423 (2017). arxiv:1702.08423 http://arxiv.org/abs/1702.08423

Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikinen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619. DOI:https://doi.org/10.1016/j.imavis.2011.07.002

Chuan Zhou and Xueyin Lin. 2005. Facial expressional image synthesis controlled by emotional parameters. *Pattern Recognition Letters* 26, 16 (2005), 2611–2627. DOI:https://doi.org/10.1016/j.patrec.2005.06.007

Y. Zhou and B. E. Shi. 2017. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 370–376. DOI:https://doi.org/10.1109/ACII.2017.8273626

Wenhui Zhu, Yiqiang Chen, Yanfeng Sun, Baocai Yin, and Dalong Jiang. 2004. SVR-based facial texture driving for realistic expression synthesis. In *Proceedings of the 3rd International Conference on Image and Graphics (ICIG'04)*. 456–459. DOI:https://doi.org/10.1109/ICIG.2004.137

Received February 2018; revised July 2018; accepted September 2018