The Promises and Perils of Automated Facial Action Coding in Studying Children’s Emotions
Aleix M Martinez
The Ohio State University

Author Note
Aleix M Martinez is with the Dept. Electrical and Computer Engineering, and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210.
The author and the research described in this paper were supported by the National Institutes of Health, grants R01-DC-014498 and R01-EY-020834, the Human Frontier Science Program, grant RGP0036/2016, and by the Center for Cognitive and Brain Sciences at The Ohio State University. The author thanks Qianli Feng, Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Shichuan Du for discussion. The Ohio State University is licensing some of the computational tools developed in the author's lab.

© 2019, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/dev0000728
Abstract

Computer vision algorithms have made tremendous advances in recent years. We now have algorithms that can detect and recognize objects, faces and even facial actions in still images and video sequences. This is wonderful news for researchers that need to code facial articulations in large datasets of images and videos, since this task is time consuming and can only be completed by expert coders, making it very expensive. The availability of computer algorithms that can automatically code facial actions in extremely large datasets also opens the door to studies in psychology and neuroscience that were not previously possible, e.g., to study the development of the production of facial expressions from infancy to adulthood within and across cultures.

Unfortunately, there is a lack of methodological understanding on how these algorithms should and should not be used, and on how to select the most appropriate algorithm for each study. This paper aims to address this gap in the literature. Specifically, we present several methodologies for use in hypothesis-based and exploratory studies, explain how to select the computer algorithms that best fit to the requirements of our experimental design, and detail how to evaluate whether the automatic annotations provided by existing algorithms are trustworthy.

Keywords: facial action coding, FACS, facial expression, emotion, computer vision, machine learning
The Promises and Perils of Automated Facial Action Coding in Studying Children’s Emotions

Facial expressions of emotion (i.e., facial configurations assumed to have emotion meaning) are believed to be important in the understanding of emotion and, thus, have played an important role in many developmental studies (e.g., Holodynski & Seeger, 2019; Leitzke & Pollak, 2016; Castro et al., 2017; Reeb-Sutherland et al., 2015; Gaspar & Esteves, 2012; Bennett et al., 2005). This hypothesis is based largely on adult recognition studies involving facial configurations described by contemporary researchers based largely on early writings of Darwin and Duchenne (Barret et al., submitted; Martinez, 2017a; Ekman, 2016). Yet relatively few studies have tested these hypotheses by examining production differences of these configurations in a wide-range of real-life social interactions in children and adults or characterized their evolution over development.

These studies have been hampered by the labor-intensiveness of manual facial behavior coding (Oster, 2003, 2006). Thankfully, in recent years, a number of commercial products have become available that purport to allow for the automated coding of both emotional and non-emotional facial configurations. Automated coding offers the promise of radically reducing the labor-costs of studying facial expression production and allowing for important scientific breakthroughs in our understanding of the relationship between facial expression and emotion and its development from infancy to adulthood. At the same time, indiscriminate use of such systems may lead to inadequate studies and inaccurate conclusions, a problem that might have already started.

It is thus urgent for developmental psychologists to understand how to use and not to use these systems and how to select the best algorithm for each study. The purpose of this paper is to inform potential users of these promises and perils. In doing so, it will provide guidelines that researchers may use to help them determine whether and how they can effectively employ such systems and which system works best on each research study. In addition, it will provide a beneath-the-hood look at how automated coding systems operate and a behind-the-scenes look at how they are developed.

Before we begin, it is imperative to understand the difference between algorithms that recognize facial articulations and those that identify prototypical facial expressions of emotion,
e.g., the prototypical facial expressions of joy, surprise, sadness, anger, disgust, fear, typically called “basic” emotions (Ekman, 2016). There is mounting evidence suggesting that these prototypical expressions are no different than other expressions of emotion (Martinez, 2017b; Martinez & Du, 2012). For example, Du et al. (2014), Du & Martinez (2015) and Srinivasan & Martinez (2019) demonstrate there are multiple facial expressions of joy, and that AU variability across subjects is relatively common. Furthermore, not every expression of happiness indicates a person is cheerful, nor does its absence indicate they are not joyful (Barrett et al., in press). For example, most people can easily fake a spontaneous (Duchenne) smile, and not everyone smiles at everything they find funny. Hence, computer systems that aim to solely recognize these prototypical expressions should be avoided. Instead, developmental psychologists should use algorithms that produce an automatic coding of facial muscle articulations. Equally important is to note that only the externalized facial articulations are observable, not the actual (internal) affective state experienced by the subject. In fact, the internal affective state may be unknowable and subject dependent (Barret et al., in press), i.e., not everyone that smiles or claims to be happy may be experiencing the same affective state.

What we can ask is whether in specific contexts, facial articulations carry affective meaning on average across the population. For example, multiple studies have evaluated the neonatal imitation of adult facial articulations and the narrowing of infant facial expressions toward those produced by caregivers (Camras, 2019; Oostenbroek et al., 2013). Such hypotheses can be studied with the help of automatic facial action coding as we discuss below. Here, it is important to keep the context fixed, since context may affect our facial articulations and interpretation (Barrett et al., in press), e.g., a person making an angry expression while telling a joke is not interpreted to mean the same as one making the same expression in a bar fight, and the interpretation of an infant’s frowning when tired or when pressed to eat its lunch is probably different too.

A related example is in the study of the communication of emotion categories versus affect in infants and toddlers. For example, Castro et al. (2017) found a clear distinction of valence in the expression of 7- to 9-year old children, but not of emotion categories. Automatic coding of facial action coding provides a mechanism to study this and related hypotheses in hundreds or even thousands of subjects in multiple cultures and contexts, which would provide a new window to the development and learning of facial communication.
Facial Articulations

There are 28 distinct facial muscle articulations that clearly fit in the above-defined framework. These are called Action Units (or AUs for short), each with a unique number between 1 and 44 (Ekman, Friesen, & Hager, 2002; Figure 1a). Four additional action units, AUs 55-58, specify head pose, and another four, AUs 61-64, denote the direction of eye gaze (Ekman & Rosenberg, 2005).

Of the 28 facial AUs, about 10 cannot be performed independently of others; for example, AU 18 (lips pucker) cannot be performed in unison with AU 28 (lip suck). That still leaves us with 18 AUs that may be co-articulated. Assuming one can move these AUs without affecting others, people can produce 262,144 facial configurations, an astronomical number. We define facial configurations as a face with any combination of AUs, while a facial expression is a facial configuration that carries some biological or sociological meaning, e.g., an emotion or a grammatical marker (Martinez, 2017a).

Infants are especially adept to move their facial muscles independently of one another; in adulthood only trained actors may be able to perform this miraculous number of facial configurations. The hypothesis is that culture modulates the expressions we perform, establishing a set of dependencies between AUs (Martinez, 2017b). This is similar to language. Infants are capable of understanding the sounds of all human languages (Werker & Tees, 1984; Gervain & Mehler, 2010), but as we grow and specialize in one or two languages, we lose the ability to produce and hear sounds used by languages other than ours. Native Spanish speakers, for example, generally add the ‘e’ sound to English words starting with an ‘s,’ thus student is pronounced /estū-d(ə)nt/ rather than /stū-d(ə)nt/. The reason for this is simple, Spanish words may start with ‘es’ but almost never with ‘s’ alone. Thus, native Spanish speakers have engraved this dependency in their speech production and breaking the pattern in adulthood requires endless practice. It appears the same is true for AUs (Martinez, 2017b; Chu et al., 2019; Camras, 2019; Oostenbroek et al., 2013). As we become more adept at communicating non-verbally with our peers, we learn AU dependencies that are difficult to break later in life.

This means that the production and perception of facial expressions evolves with age and, hence, we wish to understand this developmental process in humans. A few important, yet unanswered questions are as follows. How many of the 262,144 facial configurations do infants really produce? Of the $n$ facial configurations typically used by infants, which ones survive to
adulthood? Of the $m$ that survive to adulthood, how many are cross-cultural and how many
cultural-specific? Are the cross-cultural ones the most typically produced by infants, indicating a
biological origin? Or are they learned? Under what circumstances are these expressions produced
by infants, children, and/or adults? Is there evidence suggesting that some of these are related to
some form of affect or emotion? And so on and on.

Why have these questions not been addressed by the research community yet? The reason
is simple. To answer them, we would need to:

1. collect hundreds of thousands, or perhaps millions of hours of video of facial
   configurations in humans of all ages, from infancy to adulthood,
2. code the action units in each frame of these videos,
3. and perform a statistical analysis of the data.

Cameras are now so ubiquitous and people so eager to help science that our first point
above may be finally solvable. There are indeed privacy issues that need to be addressed, but
with proper care, availability of data should no longer be a major hurdle. We urgently need a
consortium of research teams that will collect such a dataset, and I hope funding entities will
support this effort.

Advances in statistical pattern analysis and machine learning also make the third of our
points listed above a solvable one. Machine learning is a set of computer algorithms derived to
extract useful information from data, while statistical pattern analysis identifies patterns in that
data. Older machine learning algorithms were unable to analyze large amounts of data, a problem
that is known as saturation, meaning that after a certain number of samples, the algorithms are
unable to improve their analysis. But recent algorithms, especially deep learning (Goodfellow et
al., 2016), are capable of working with very large datasets, what is now commonly term big data
(Martinez, 2017a).

But how about point 2 above? How can we code action units in thousands or even
millions of hours of video? As a simple example, consider the analysis of a full year of video for
each of 100 subjects. That is 876,000 hours of video or 98 billion frames (assuming 30 frames-
per-second).

If we are ever to complete a study of this magnitude, our only hope is to have computer
vision algorithms that do the coding of AUs in each video sequence, video frame or still image
automatically, with minimal or no human intervention. Computer vision is an area of artificial
intelligence devoted to the design of algorithms capable of automatically analyzing images and videos; for example, identifying AUs in images of facial configurations. Computer vision algorithms capable of coding the action units of faces in images and videos are already available, Figure 1b. How do these algorithms achieve this feat? How and under which conditions can developmental psychologist use them? Are the results of these algorithms trustworthy?

This paper provides concrete answers to these and related questions. I summarize the different types of computer vision algorithms that are available for coding AUs. I also put forward a proposal of best practices for those wanting to use these algorithms. Therefore, this paper will be primarily useful to researchers who want to use these automated systems to answer some of the fundamental scientific questions listed above. I will list the dos and don’ts and explain how to design experiments to maximize the likelihood of success and reproducibility. Note that the scientific questions I enumerated earlier may be answered using exploratory experiments and, hence, I will describe how these should be properly conducted. But, I will also describe how these computer vision algorithms can be added to a hypothesis-based experiment, since these tend to be preferred by researchers.

I cannot emphasize enough how important it is to use well-designed methodologies and follow the best practices outlined below when using computer vision algorithms. Computer vision algorithms are no substitute for well-designed experiments, nor are they good enough to fully substitute human experts, at least not yet. There is much these algorithms can help us achieve, but only if we take care of the important methodological details and limitations described below.

Finally, some of the computer algorithms described below are available from companies. These require little effort from the researcher, but, as we explain below, these may not be the most appropriate algorithms to study the scientific question of interest. Thus, it may be necessary to use the algorithms provided by some computer vision research groups. Using these may require to hire a computer scientist (ideally a computer vision specialist) to run the experiments for us. The details provided below will help you decide whether this is necessary.

The research conducted in my research lab was approved by the Office of Responsible Research Practices at The Ohio State University, and subjects provided written consent; study title “Face Recognition: Data collection, recognition of identity and expression,” study number 2002B0258.
Proper Experimental Design

To properly incorporate automatic AU coding in our experiments, we must first carefully define what we wish to accomplish. This is important, because some computer vision algorithms are best suited for some specific tasks more than others. Also, existing algorithms may or may not adapt to our experimental requirements, or may lead to incorrect predictions or irreproducibility (Poursabzi-Sangdeh et al., 2018).

Look at the taxonomy of experimental settings given in Figure 2. We start with a well-defined experiment. A properly defined experiment will specify which type of data collection we intend to assemble, or which one was made available to us. Generally speaking, there are three conditions, which I have termed: i. idealized conditions, ii. in-lab-like conditions, and iii. in the wild. The first of these groups includes images and videos filmed indoors under good, non-changing illumination, with a frontal view of the subject’s face, and no occlusions. The second group includes images and videos for which illumination may vary but are almost always well-lit; may be filmed indoors or outdoors; if not frontal, faces appear at an angle that allows humans to distinguish the desirable AUs – this is usually between -20° and +20° in each of the three axes of rotation; have minor occlusions that do not block the AUs of interest. The third group are images collected under completely unconstrained conditions, which is typically termed “in the wild,” analogous to studying biological systems in the wild, rather than in the lab. While we generally prefer to study expressions in the wild, this is obviously the most challenging setting for computer vision algorithms.

Idealized conditions

As the reader may suspect, computer vision algorithms perform best with still images and videos in the first group – images and videos collected under ideal conditions. However, collecting and curating these images and videos requires a big effort by the experimenter. For example, it is extremely unlikely one will always obtain a frontal view of every filmed facial configuration. If that is the requirement of our computer vision algorithm, then human annotators will have to sweep through every still image or video frame to determine which can be used in our study and which cannot. This may be problematic for a variety of reasons. First, it may be a monumental task that takes months or even years to complete. It is of course much simpler and preferred to having to manually annotate AUs, but it is still a time-consuming task. Second, studying only frontal faces may eliminate important fundamental variables of interest in our
study. Maybe head pose is a much better determinant of what we wish to study than AUs, or maybe some AUs of interest are only produced when the head is tilted in specific directions away from the camera; additionally, head pose is known to influence our reading of a face (Witkower & Tracy, in press; Lyons et al., 2000). Third, the need for non-occluded, frontal faces may require placing the camera in a location that prevents natural, spontaneous expressions from occurring. And so on. The important conclusion is that computer vision algorithms can accurately annotate our images and videos if we have selected the images and videos carefully to almost exclusively include frontal, non-occluded faces that are well illuminated with a non-changing light source, but these images and videos may be unfit to answer our scientific questions.

If your research does fit within this first group, the next question you need to ask is whether you are interested in analyzing still images or video sequences, Figure 2. You will use images when you wish to study facial configurations at specific time points, and you will use video when changes over time is a variable of interest. If your experiment requires that you study facial configurations at specific time points, you need to ask one final question: do the images in these time periods correspond to the apex of the facial configuration? If the answer is yes, then, there exist several computer vision algorithms that you will be able to adapt to your experiment. Of course, if the answer is yes, this means that you or your colleagues have been very careful when collecting your images to make sure that they correspond to the apex of each facial configuration, or you have curated the images to select those that show only the apex. In these cases, there are several computer vision algorithms that may be adapted to your experiment, as discussed later in this paper.

If, on the other hand, you do not know whether the images correspond to the apex or not, then no computer vision exists that can find the apex for you. Most non-computer vision experts will be surprised by this. Detecting the apex of a facial configuration seems such a simple problem. If a video sequence is available, all that needs to be done is to identify the frame where the AUs are at maximum activation. Note that AUs can be activated at multiple intensities. For example, AU 12, which indicates the outer pulling of the corners of the lips (i.e., the corners of the lips move away from the center of the face), can be activated at different intensities (i.e., increasing the distance from the center of the face as a function of intensity). Thus, one may think that detecting the apex simplifies to identifying the point of maximum extension of the
AUs. There are a few problems with this hypothesis though. First, current computer vision algorithms are very good at detecting AUs in the idealized imaging conditions described above, but not as good at detecting the intensity of activation. Some algorithms do detect the intensity but not with enough precision to allow us to accurately specify the apex of the facial configuration. Second, the maximum extension of each AU is subject dependent. What may be maximum activation for me, may be only half active for you. Facial muscle differences between people constitute a barrier for a mathematical definition of apex. Third, and arguably most important, AUs come and go rapidly, generating a large number of activation peaks, but not all of these peaks define the apex of a meaningful expression. Actually, most of these activation peaks correspond to transitional facial movements unrelated to any variable of interest to the researcher. Consider, as a simple example, a subject yawning or sneezing. How will the computer vision system know this is not a facial expression as defined above?

To further clarify the last point in the preceding paragraph, consider the following. First assume the data has been carefully curated to include only video sequences with a single facial expression. Here, detecting the apex of an expression is relatively easy. Now consider a video of a long conversation between two subjects. This includes a barrage of facial configurations. Some include coding of expressions in isolation (e.g., a surprised reaction to a comment) which are easy to detect. But one expression may overlap another, e.g., a subject starts to express surprise to the comment of another person when she realizes that it was a joke and starts to laugh before the surprise expression reaches the apex. How would you code for this? Is the maximum extension of the surprised expression to be consider the new apex? Even if that does not reflect the maximum intensity of the AUs in those frames of the video sequence? If so, do we impose a minimum intensity of the AUs to consider it an apex? This opens a fundamental, recursive problem in our experimental design: In most cases, we are interested in identifying facial expressions unknown to us and, hence, we wish to not constraint what defines the apex of that expression. But without such a definition, the number of meaningless facial configurations will be unmanageable, e.g., are the AUs of a sneeze relevant? How about those due to breathing, babbling, speech, etc.? At present the only solution is to have a human expert in-the-loop who defines what is meant by apex, evaluates the outcomes of the computer vision algorithm carefully, or both.
A human-in-the-loop means that the computer vision system and the human work in unison at each step in the experiment. For example, after curating our images as much as possible, we still do not know which images define the apex of a potentially meaningful facial configuration and which do not. To solve this, we first use an appropriate computer vision system to automatically annotate the presence of AUs. AU patterns (sequential or otherwise) that repeat multiple times over time and across subjects are selected. This selection needs to be done carefully and must be based on some grounded assumptions; for instance, what is the minimum number of AUs we are interested in and why? And, how many times does a pattern of AUs need to occur to be significant? These and other questions must be carefully answered by the research team, not by the algorithm. This exercise will give the research team a set of potentially interesting facial configurations. Careful analysis of these facial configurations must follow. Do any of them correspond to autonomous or semi-autonomous body movements (e.g., a sneeze) or babbling or anything else we are not interested in? If so, we need to identify the properties of these facial configurations and ask the software to redo the analysis with these additional constraints added. This process must be repeated until we identify the facial expressions we were looking for in our study.

The problems enumerated above (e.g., specifying the apex or the minimum number of AUs) can be alleviated when we perform a hypothesis-based experiment. If we wish to test an accepted or well-reasoned hypothesis, all we need to do is to determine if the process described in the preceding paragraph yields the expected facial configurations. For example, a famous hypothesis is that humans of all cultures share six facial expressions – joy, surprise, sadness, anger, disgust and fear – sometimes called “basic” emotions (Ekman, 2016; Jack et al. 2012). If we studied millions of still images of facial configurations across cultures would we identify these expressions and these expressions only as universal? Or would we identify a much larger number of expressions, as hypothesize by other authors (Du et al., 2014)? We recently used the above experimental design to test this hypothesis (Srinivasan & Martinez, 2019), and identify a much larger number of cross-cultural expressions, 35 to be precise, demonstrating that people regularly use many more facial expressions of emotion than previously believed. This study included a carefully collected dataset of about 7.2 million images and 10,000 hours of video which were collected online using web search tools in 30 different countries.
Let us now go back to Figure 2 and consider the case where we wish to automatically annotate AUs in video sequences collected in idealized conditions. We see that this becomes a solved problem again. However, this statement needs to be qualified. What is meant here is that computer vision algorithms will be able to automatically annotate AUs accurately (Benitez-Quiroz et al., 2019), not that this will give us any meaningful scientific analysis. If we wish to identify expressions in this newly annotated video set, we will need to use the approach defined above for still images all over again. A human will need to carefully evaluate the results of the algorithm to identify any expression of interest.

**In-lab-like conditions**

Most likely, your still images and video sequences will be collected in somewhat controlled conditions but not idealized ones; maybe indoors, with some 3D head rotation and minor occlusions. Current computer vision algorithm can mostly deal with these imaging conditions and, hence, the procedures to follow do not deviate much from the ones already described in the preceding section.

If you are interested to work with images that have been curated to show the apex of a facial articulation, and/or wish to work with video, then, there are several computer vision algorithms you may use to automatically annotate AUs. As shown in Figure 2, still images are easier because you have already indicated they display the apex of the facial configuration, you will only need to select the right computer vision algorithm, a topic we will discuss in detail in later sections of this paper. If you can show that the selected algorithm works as expected on your database of images, you will be able to trust your annotations.

As for video analysis, you will need to check that the analyses given by the computer vision algorithm provide reliable results. The amount of human involvement will depend on the task you are interested to solve. As stated above, if your goal is to detect the apex of all facial configurations, then you will need to provide additional information to the algorithm to define what this really means. Make sure your model or assumptions are based on well accepted theories and/or experimental results, or you run the risk of falling into a circular trap. As an example, consider the following problem: We are interested in identifying the number of facial expressions infants produce. To address this, we use a computer vision algorithm to analyze thousands of hours of videos of infants interacting with their parents. We define a facial expression as the point in which the number of AUs is maximal for every small interval of $t$
seconds. After completion of our study, we conclude infants produce \( q \) facial expressions. But closer inspection shows that some expressions have been missed, because some important expressions overlap with others yielding a monotonically increasing number of AUs (e.g., a frowned started before the conclusion of a smile), but our definition only allowed us to detect the one with the larger number of AUs in each interval of \( t \) seconds (e.g., a smiley frowned that had no intentional meaning). Computer vision and machine learning algorithms will compute what we define, not necessarily what is needed. But if we do not have a mathematical definition of what we wish to uncover, we cannot ask the computer vision algorithm to solve it. In fact, many scientific studies are performed to identify that definition, but, in these cases, the intrinsic definition coded in the algorithm will be the one we identify, whether we are aware of it or not. Similarly, in hypothesis-based experiments we will most likely find results that support our hypothesis if that definition is what was given to the computer vision algorithm. One solution to this problem is to run a permutation test, as described later in this paper.

As shown above, if we wish to analyze still images of faces that may or may not display the apex of a facial configuration, then the problem can only be solved with a human in the loop. One way is by specifying what AUs and intensities we are interested in and whether there is a requirement on the number of AUs per expression. Another solution mentioned above is to ask a human expert coder to verify that the automatic annotations provided by the computer vision system are accurate (Srinivasan & Martinez, 2019).

**Images and videos in the wild**

In some instances, we may wish to analyze images and videos collected in the real world, under completely unconstrained conditions. Here, images and videos may be low quality, have large variations in illumination, pose, ethnicity, skin color, have major occlusions, etc. Even in the lab, developmental scientists working with children (who often have difficulty remaining relatively still) may find large variations in these factors that will pose challenges. Can we use computer vision algorithms to automatically code AUs in these still images and video sequences? While extra care will need to be taken when doing this, the answer is yes, at least in some cases.

As we will detail in the section to follow, assuming our still images have decent quality, the major AUs of interest are not occluded, and images represent the apex of facial configurations, computer vision algorithms already exist that can provide a reasonable (useful)
annotation of AUs (Srinivasan & Martinez, 2019; Benitez-Quiroz et al., 2016). In these cases, we will still need to verify the results manually, but that is a much-preferred task over that of manually providing the annotations ourselves.

However, when the still images may or may not represent the apex, or when we use video sequences, the problem can only be solved with a human-in-the-loop, Figure 2. As in the above, the amount of human work will depend on the goal of the project, but, at a minimum, a human expert will have to carefully monitor and evaluate the performance of the algorithm at every step of our study.

**Automatic Coding of Action Units**

**Spatial versus dynamic representation**

As mentioned above, one of the most important decisions we need to make when selecting a computer vision system to automatically code facial action units is whether we are interested in their functional change of activation over time or on the discrete activation at several time points. The former requires the analysis of video, while the latter can be performed in video and images. Let us first explain what the basic differences between these two analyses are.

**Recognition of action units in still images and video frames**

This is the most typical analysis seen in scientific studies to date. The goal is to identify which AUs are present in each of a set of available still images or video frames (Martinez, 2017a; Du & Martinez, 2012). If we are given a set of images, our goal is to select an algorithm that can tell us which (if any) AUs are active in each of the faces that appear in the images. If we are given a video sequence instead, then the algorithm must list the active AUs in each of the frames of the video sequence. In some instances, we may also be interested in an algorithm that can specify the intensity of activation of each AU. The standard way to specify intensity is to categorize each AU into one of five values (Ekman & Rosenberg, 2005), namely: a (meaning there is only a trace of the presence of this AU), b (indicating a slight activation of the AU), c (meaning the AU is clearly marked), d (specifying the activation of the AU is extreme), or e (indicating the activation of the AU is maximal). Figure 3a shows an example.

**Functional representation**

Another way to study facial actions is to uncover the underlying function of muscle articulations, Figure 3b. While the methods described in the previous paragraph provide a
qualitative analysis of the presence of AUs in each image or frame of a video sequence, the methods used here yield a quantitative analysis of the activation change over time (Simon et al., 2010). This means that each AU is defined as a function (i.e., a curve) over time, rather than a set of discrete letters as above, Figure 3. As shown in this figure, one may be interested in the intensity of activation $f_i$ or some other variable, e.g., co-articulation of AUs, frequency or probability of activation, etc.

Once we have determined which variables are of interest to us and whether a qualitative or a quantitative measurement is needed, we ought to identify a computer vision system that can provide the values of these variables by analyzing a large dataset of images or videos. This means we need to understand which systems are available and what they can and cannot do, and what is the degree of accuracy of their analyses.

**Computer Vision Methods**

The computer vision algorithms that have been designed to label AUs in still images and video sequences use either computer vision approaches or machine learning techniques or a combination of the two. Figure 4 summarizes the main techniques used by the algorithms. The professional computer systems made available by companies as well as those made available by researchers fit within one of these groups. It is important to understand the approach used by the selected algorithm for two reasons. First, we want to select a method that has been demonstrated to perform well under similar imaging conditions to those of our data. To do this, we need to know the approach used by the selected algorithm by reviewing the papers or reports were the system is defined. Second, this same report should provide a description of the images used to evaluate the algorithm. Use this to determine whether this is a good fit. After this, we will need to test whether the selected algorithm works on our dataset. If not, we should generally avoid algorithms that use the same approach and move on to algorithms that employ distinct strategies.

Let us briefly summarize these distinct approaches.

**a) Template matching.** Template matching is a classical approach in computer vision. As its name indicates, given a template, the goal is to find whether it is present in an image and, if so, where (Martinez & Kak, 2001). When applied to automatic detection of AUs, we first need to generate a template of each AU. For example, one can define a window $w_i$ of $p \times q$ pixels centered at the place of articulation of AU $i$ on a number of sample images of faces with that AU active. Statistics are then extracted from these sample
windows, e.g., the mean, standard deviations, covariances, etc. If we use the mean and covariance matrix, we define the image variability of that AU template using a Gaussian distribution. Given a test image, we extract the window \( w_t \) of \( p \times q \) pixels centered at the location of AU \( i \) and calculate the distance to the computed Normal distribution, e.g., using the Mahalanobis distance or Bayes error (Zhu & Martinez, 2006; Hamsici & Martinez, 2007). If that distance is below a threshold, we say that AU \( i \) was detected in the image. We can make this system better by using a mixture of Gaussians instead (Martinez & Vitria, 2001). Alternatively, we can compute the distance to the subspace of principal or independent components, given by Principal Components Analysis (PCA) and Independent Components Analysis (ICA), respectively. The PCA and ICA representations are linear, meaning the statistical model that represent AUs is given by a linear equation (Draper et al., 2003). Nonlinear manifolds allow us to define more flexible models. This is achieved by either changing the metric of the feature space representing \( w_t \), a technique called kernel mapping (You et al., 2011), or with nonlinear regression (Rivera & Martinez, 2012).

b) Optical flow. Template matching can also be used to determine the movement of fiducial points. This is called optical flow and is defined as the apparent motion of the brightness pattern of a set of images (Baker et al., 2011). However, while the template matching method described above is used to detect the presence of an AU, here its purpose is to uncover the movement of a fiducial point across a number of video frames or images, e.g., the outer pulling of the corners of the mouth of AU 12. This can be readily computed from video sequences that start at a neutral face followed by the activation of a number of AUs (Lien et al., 1998; Donato et al., 1999; Martinez, 2003a; Liu et al., 2016). When only an image is available, we ought to compare it to a neutral face, ideally of the same individual, but, if unavailable, of a norm facial identity (Martinez, 2003b; Du & Martinez, 2012). Figure 5 shows some examples of the optical flow estimated on a single image and the mean neutral face of a large number of individuals. As can be seen in this figure, optical flow provides a direct measure of the perception of apparent movement of the facial muscles, which can be used to identify AUs.
c) **Image filters (Gabor, Wavelets).** A window of $p \times q$ pixels, called a kernel, is centered at the location of each AU and convolved\(^1\) with that local region of the image. This process typically yields distinct results when the AU is active than when it is not, allowing computer vision algorithms to identify its presence or absence. Kernels that have been shown to yield this distinction are variants of the Gabor kernel and wavelets (Lyons et al., 1999; Tian et al., 2002; Yang et al., 2007; Savran et al., 2012). This approach is usually applied on several pixels around the center point of each AU to add robustness to its detection. One of the arguments for using Gabor filters is their resemblance to the computations executed by our own early visual cortex (Martinez & Du, 2012), which may aid in the classification of AUs thought to occur in a nearby brain region called the posterior Superior Temporal Sulcus (pSTS) (Srinivasan et al., 2016).

d) **2D and 3D shape analysis.** Above we defined a way to detect facial movements with optical flow. Another popular approach is to use Procrustes analysis (Hamsici & Martinez, 2008, 2009a, Sun et al., 2008; Garg et al., 2013; Jin & Tan, 2017). Procrustes analysis is an algorithm to align a set of fiducial points (e.g., corners of the mouth, center of the eyes) across multiple images. This registration allows the algorithm to compute the deformation of the shape of the face using a statistical model, as, for example, Principal Components Analysis (PCA) (Martinez & Kak, 2001; Martinez & Zhu, 2005). Using PCA implies we compute the mean and covariance matrix of the deformation of the face, meaning the facial expression is modeled using a Normal distribution (Todorov et al., 2016). This can be extended to more complex distributions by changing the norm of the space (Hamsici & Martinez, 2009b), which also allows us to recover the 3D shape of the facial expression, as shown in Figure 6a, as well as other transformation functions (Agudo & Moreno-Noguer, 2018a). An alternative approach is structure from motion, which estimates the movement of face (scales, translates, and rotates with respect to the camera) as well as its 3D shape (Jia & Martinez, 2009; Gotardo & Martinez, 2011a,b; Agudo et al., 2014; Agudo & Moreno-Noguer, 2018a). As with Procrustes analysis, a change in the metric (called a kernel mapping) is typically used to improve the results and, in addition, it allows us to recover the 3D shape of the face (Hamsici et al., 2012; Gotardo & Martinez, 2011c). Figure 6b shows an example sequence. An alternative to

---

\(^1\) A convolution is given by adding the elements of the image within a determined window, weighted by the kernel.
kernel maps is the use of sparse representations, which simplifies the number of unknowns to be solved to yield robust results (Li et al., 2015). And, finally, some models combine the formulation of Procrustes analysis and structure-from-motion to compute the shape of the face (Lee et al., 2013), while others use deep learning (Zhao et al., 2018; Albiero et al., 2018; Chang et al., 2018).

e) **Isoluminant color and shading.** The shading of a face is given by the luminance, or quantity of light per unit area at each point on the surface of the skin; this is 1-dimensional and can be readily computed by mapping a color image into grayscale (i.e., from three color channels to one). Isoluminant color is what is left in the image once the luminance has been factored out. Thus, isoluminant color is 2-dimensional. We believe the human visual system uses two opponent color channels – yellow-blue and red-green – to represent images and objects (Gegenfurtner, 2003). It has been recently shown (Benitez-Quiroz et al., 2018) that facial color in this isoluminant color space changes as a function of the emotion experienced by the expresser. The assumption is that hormonal changes have an effect on facial blood flow and/or composition that is visible through color variations on the surfaces of the skin of the face. This information can be combined with shading cues, which defines the 3D shape of the face, to detect AUs with greatest accuracy than ever before (Benitez-Quiroz et al., 2019).

**Machine Learning Methods**

The computer vision systems defined above are formulated based on our understanding of the physics of the world (e.g., light, geometry) and existing computational models of the human visual system (Martinez, 2017b; Martinez & Du, 2012). Another solution is to learn the representation that is best suited for a specific dataset. This is the goal of machine learning.

f) **Classifiers.** Deep feature representations have become commonplace (Benitez-Quiroz et al., 2017b; Bai et al., 2018; Pons & Masip, 2018; Corneanu et al., 2018). Given a large dataset of images of facial configurations, we use a deep neural network and train it to identify action units. A deep neural network is composed of a number of layers generally represented as a directed acyclic graph (Goodfellow et al., 2016). The outputs in the last layer correspond to the classification of AUs, and the previous layers correspond to the so-called “deep features.” We can use these deep features in lieu of the computer vision features defined above. While this approach has yielded top results in other computer
vision problems, computer vision features yield equally good and, in many cases, better results than these deep representations (Benitez-Quiroz et al., 2017a). Discriminant analysis, a statistical pattern recognition approach (Hamsici & Martinez, 2008; Zhu & Martinez, 2006; Deng et al., 2018; Wan et al., 2018), is also used to uncover the best predictors of AUs (Benitez-Quiroz et al., 2016), while other algorithms use Support Vector Machines (SVMs) over the computer vision or deep representations described above (Bartlett et al., 2005; Kotsia & Pitas, 2007; Zhang et al. 2014; Du & Martinez, 2014; Girard et al. 2015).

g) **Unsupervised methods.** Machine learning algorithms are tasked to find the functional mapping $y_i = f(x_i)$, where $x$ is the input feature vector (which may define one or more of the computer vision features given in a-e, or be a set of deep features as explained in f) and $y$ is the desirable output, e.g., $y_i = (y_{i1}, ..., y_{ip})^T$, with $y_{ij} = \{-1, +1\}$ indicating the AU $j$ is present (+1) or not present (-1); alternatively $y_{ij}$ may define the intensity of activation of AU $j$. The machine learning algorithms described above use a labeled training set, $\mathcal{Y} = \{x_i, y_i\}_{i=1}^d$, to find a possible mapping $f(.)$, and $d$ is the number of training sample pairs. The algorithms using this labelled dataset are called supervised methods, because the task of finding $f(.)$ is determined (supervised) by $\mathcal{Y}$. The problem with this approach is that a human expert must provide a large training set $\mathcal{Y}$, and, as we know, manually annotating AUs in a large number of images or video frames is costly; that is why we wish to use automated computer vision systems instead. Therefore, the main goal in modern computer vision and machine learning is to define algorithms that can learn from a large set of unlabeled data, $\mathcal{X} = \{x_i\}_{i=1}^d$. This is called unsupervised learning, because the labels ($y_i$) are not given. As of this writing, unsupervised learning of action units is still an open area of research. Recently, Zhao et al. (2018) have derived an algorithm that can learn from a large set of unlabeled internet face images. The key idea is to group image as a function of image feature similarity and image description similarity using techniques from graph theory. Some other recent methods (Wiles et al., 2018) are not specifically defined to detect AUs in faces, but may be adapted to achieve this goal in the near future.

h) **Generative models.** If the functional mapping $f(.)$ is given by a probabilistic model defined by an underlying but unknown density function, we can use probabilistic
algorithms to estimate it. The most classical approach to density estimation is mixture models (Reynolds, 2015), with a long tradition in modeling a variety of visual stimuli (Martinez & Vitria, 2001). Most algorithms model the activation of AUs using a mixture of Gaussian (Song et al., 2015), with variants using a mixture of PCs and ICs (Draper et al., 2003). When adding time to these models (i.e., in video analysis), we have a Hidden Markov Model, which has also been successfully used to model facial expressions (Corneanu et al., 2016; Cohen et al., 2003; Martinez, 1999). Deep learning methods can also be used to estimate the underlying distribution, with the most tested approach being Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Pumerola et al. (2018) have used this approach to learn the underlying distribution of the image changes of every AU. This means that given any arbitrary image, this algorithm can edit it to add or subtract any AU from the image. As seen in Figure 7, the results are so convincing that it easily tricks human subjects in believing that the generated images are in fact real. Thus, this approach can be used to detect AUs in images as well as to generate new stimuli for our experiments. A variety of applications of this approach and extensions are already underway (Romero et al., 2018; Vielzeuf et al., 2018).

**Evaluations**

To date, the computer vision and machine learning algorithms described above have been mostly evaluated on data collected in the laboratory, under constrained conditions, and only a handful of algorithms have been tested with still images filmed in unconstrained conditions outside the lab. Here, constrained conditions may refer to illumination, pose and other image collection mechanisms or to the restricted way in which subjects are asked to behave; in general, people do not act naturally in the lab, while illumination and pose are at least somewhat constrained.

It is important we understand under which conditions each algorithm has been tested. When selecting one of the computer vision or machine learning algorithms described in the preceding section, we need to check how these were tested and evaluated to make sure these can be used and were evaluated with the same type of images and videos we wish to automatically analyze.

There are three types of data on which algorithms are typically evaluated.
i. *Posed expressions:* These are still images or video frames of typically hypothesized facial expressions. Subjects are asked to pose the expressions by either imitating the expression in an image, following a cue (e.g., smile, frown), or giving subjects a situation and asking them to produce the expression that would be expected in it. Examples are the CK+ and the Compound Emotions datasets (Lucey et al., 2010; Du et al., 2014).

ii. *Spontaneous facial configurations:* Here videos and images are collected while subjects watch a video or interact with another person, yielding several spontaneous facial configurations. Examples are DISFA (Denver Intensity of Spontaneous Facial Action) and Shoulder Pain datasets (Lucey et al., 2011; Mavadati et al., 2013).

iii. *Images and videos in the wild:* These are images and videos collected outside the lab, in completely unconstrained environments/conditions (Martinez, 2017a). The term “in the wild” refers to the fact that they are collected outside controlled, in-lab conditions. The largest dataset is called EmotioNet (Benitez-Quiroz et al., 2016) which includes 1 million images; the recent extension of Srinivasan & Martinez (2019) contains over 7 million images and 10,000 of video with more than 1 billion frames.

Obviously, when selecting a computer vision or machine learning algorithm to automatically annotate AUs, one must make sure that it has been tested on similar conditions to those of our data. For example, if your experiment only includes posed expressions, has the algorithm you wish to use been extensively tested using well-documented datasets of posed expressions? If not, then you either need to test it yourself, or you ought to find a different algorithm.

Let us assume you have now selected an algorithm that has been extensively tested on images and/or videos collected under the same imaging conditions as those of your data. And, let us further assume that these studies prove the selected algorithm performs wonderfully on those images and/or video sequences. We can now use the selected algorithm with confidence, right? Unfortunately, the answer is *no.* Why not? Because, most likely, these tests have been run by the same team that designed the experiment and the algorithm might be overfitted to the testing data. Let me explain what that means. In computer vision and machine learning we typically divide our dataset into two subsets. The first is used to train our algorithm, e.g., to identify the parameters of the algorithm that make it work well. Then, the tuned algorithm is run on the testing data, yielding the results that are to be expected when using a similar, yet independent
dataset. The problem is that the people who design the algorithm we wish to use had access to both, the training and testing data, during development, and, most likely, they modified their algorithm until it worked on both the training and testing datasets. That is, they overfitted their algorithm to the training and testing data. This means their testing results may not be a good representation of what you might expect to see on truly independent, previously unseen data.

How can we solve this problem? We have three options. One is to manually annotate AUs in a number of images or videos of our dataset, use the selected algorithm, and compute how well it does on it. The more annotations we use, the better. A second option is to use the selected algorithm to annotate our data, randomly select a number of images or video frames (say, 5% of them), and manually check the accuracy of these annotations. The third option is to find a number of annotated datasets that were not used by the developers of the selected algorithm and use these to check how well the algorithm performs on these novel datasets.

When evaluating an algorithm or looking at evaluations performed by others, do not pay much attention to the accuracy of the algorithm. Accuracy is defined as the number of correctly labeled images, divided by the total number of images used in the test. The problem is that most images do not have AU $i$ present, and a simple algorithm which always says AU $i$ is not present would have a very high accuracy but would be useless. As an example, consider AU 4. Imagine AU 4 appears in .2% of the images in a database of 1 million samples. If our algorithm says that AU 4 is not present in any of these 1 million images, its accuracy would be $\frac{998,000}{1,000,000} = .998$, i.e., the accuracy of this algorithm is 99.8%, even though the algorithm is unable to code AU 4 at all.

To address this issue, we need to compute the precision and recall of the algorithm. Precision measures the fraction of selected images with AU $i$ that are correctly classified, while recall (also called sensitivity) is the fraction of detected images with AU $i$ over all the images with AU $i$. These two measures are typically combined in a single value called $F_1$-score, $F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. $F_1$ takes values between 0 and 1, with 0 indicating the algorithm is useless at the task and 1 designating perfect performance. In our example above, $F_1 = 0$, even though accuracy= .998.

All the above will be necessary unless the authors of the algorithm have validated their algorithm on a large and truly independent dataset they had no access to. This typically means the authors have participated in a challenge or competition, where the testing data was
sequestered and, thus, not available to the team designing the algorithm. The most extensive one is the EmotioNet challenge (Benitez-Quiroz et al., 2017a), but this challenge exclusively uses images in the wild. Thus, it is generally highly recommended to test the selected algorithm on your data before you proceed; some algorithms perform really well on some datasets and really poorly on others. Having a computer vision expert in your team who can perform these evaluations is also advisable. Additionally, whenever you need to evaluate a computer vision algorithm on your data, you will also need to add a certified FACS coder on your team (e.g., to manually annotate a subset of the data for reliability purposes).

It is important to note that the more your data deviates from that of previous tests, the more likely it is for the selected algorithm to fail. For example, to the author knowledge, the highest $F_1$-scores on posed and spontaneous expressions are those achieved by the algorithm of Benitez-Quiroz et al. (2016), with $F_1 > .94$ when testing on CK+, DISFA and Shoulder Pain, which is as good as human annotations (Girard et al., 2015). However, these results were obtained by training the algorithm on a subset of each of these databases and testing it on an independent subset of the same dataset, a method called cross-validation. $F_1$-scores drop to about .6 when training on some of these datasets and testing on very different databases. This is because the imaging conditions in these datasets are extremely different, making the classification of AUs database specific. One solution to this problem is to retrain these algorithms with a portion of your database. For this though, you will need to manually annotate a portion of your data, which is time consuming.

**EmotioNet Challenge**

The most challenging problem is, of course, the detection of AUs in the wild, where the algorithm needs to adapt to any possible imaging condition. The only largescale test that assesses computer vision algorithms in these challenging conditions is the EmotioNet Challenge. Thus far, there have been two challenges, a first one in 2017 and a second in 2018. Of the dozens of participants that registered, only 10 have completed the challenge. The top $F_1$-scores are about .64 on the moderately difficult set and about .56 on the most difficult one (Benitez-Quiroz et al., 2017a, 2017b). These results improve when the facial color features associated to emotion are also considered (Benitez-Quiroz et al., 2019).

---

2 See also the results in the EmotioNet website: [http://cbcsl.ece.osio-state.edu/EmotionNetChallenge/index.html](http://cbcsl.ece.osio-state.edu/EmotionNetChallenge/index.html)
A clear limitation of AU detection in the wild is in pose invariance (Benitez-Quiroz et al., 2017a); that is, how can we recognize AUs when the face is not observed frontally? One solution is to recover the 3D shape, shading and discriminant colors of the face from a single 2D image (Zhao et al., 2016). This is an ill-posed problem, meaning that for any 2D image of a face there is an infinite number of possible 3D faces that could have generated this 2D observation. A small number of algorithms have recently solved this problem by learning the mapping function between 3D and 2D images of faces (Zhao et al., 2018; Zhao et al., 2016), with extensions and variants of these algorithms improving over previous results (Tome et al., 2017; Jourabloo et al., 2017; Rad et al., 2018).

Another problem with the above algorithms is the intrinsic biases of the databases used to learn to discriminate between AU present versus not present. As we saw above, computer vision algorithms do not perform as well when our training dataset is not a good representation of what we will be using in testing. A major issue is that most databases used to train these systems do not have a large number of images of certain ethnicities and races. Hence, algorithms trained with these databases provide subpar performance on the poorly represented groups (e.g., black subjects) (Buolamwini & Gebru, 2018). It is imperative to test your system on the demographics you will be using it with, before you decide whether that algorithm will perform as expected.

Exploratory and Hypothesis-based Designs

Let us see how we can use the information detailed above to design our experiments. First of all, we must decide whether we will perform an exploratory or a hypothesis-based study. Since researchers (and funders) typically prefer hypothesis-based studies, let us define those first.

Hypothesis-based

Knowledge is the pillar on which research rests. Typically, we will want to design an experiment to test whether an established model or hypothesis is true or not; in other cases, we may wish to test a novel hypothesis. To this end, we need to design an experiment that challenges our hypothesis. After careful thought, we determined that an analysis of facial action units in a large number of images of videos is necessary, and hope we can complete this analysis automatically, using a computer vision system. How should we proceed?

First, we ought to know whether such a system is available. Using Figure 2, we can easily determine how we can proceed and how much care one needs to take when collecting and
curating the database of face images/videos to be used in our study. Above we provided a detailed explanation of Figure 2, which we can now use to design our experiment.

Second, we need to select a computer vision algorithm from those listed in Figure 4. This selection needs to be directed by the needs of our study and the design we have already defined. If we are to analyze still images showing the apex of an expression, then we will select one of the algorithms specifically designed to work with these. If the images sometimes do and sometimes do not show the apex of an expression, then, we will need to use an algorithm that can detect which images correspond to an apex and which ones do not by providing a specific definition of what we call the apex; e.g., we may define apex as the frame of a video sequence with the maximal number of AUs, or as the point that each AU is at maximum activation, or at the point when the AU activation first increases and then decreases by a specified amount (i.e., a threshold), etc. This definition will be part of our hypothesis. On the other hand, if we wish to analyze the temporal information of AUs, then we need to select one of the algorithms that can provide a quantitative analysis over a video sequence, rather than a qualitative analysis on images. Similarly, if we are interested in the intensity of activation of an AU, we need to determine if we wish to recover a category (i.e., a set of levels of intensity), or a continuous value, and then select the appropriate algorithm.

Once we have selected our algorithm, we will need to test it on our data, as described above, to make sure it will yield an accurate analysis of our data. We are now ready to use the selected algorithm to test our hypothesis.

When testing established hypotheses, we may not want to stop here. Once we have used the selected algorithm to evaluate our hypothesis, we can modify it to accommodate the new results. This will give us a new hypothesis, which can be retested on a different dataset of images or videos and using the same approach described above. Alternatively, and preferably, we can test the new hypothesis using behavioral or imaging studies. We can repeat this process until our hypothesis has been modified to justify the new data. In fact, this is one of the main advantages of using computational analysis in big data – it allows us to modify, extend and tune currently accepted hypotheses, which will then serve as the seed of novel scientific studies (Martinez, 2017a). For example, Izard, Dougherty, & Hembree (1983) identify multiple expressions in infants, and Du et al. (2014) hypothesized the existence of compound facial expressions of emotion, like happily surprised and happily disgusted in adults. Du et al. then went a step further
by presenting a computational analysis that supported their hypothesis. Then, in Du & Martinez (2015), we tested this novel hypothesis by identifying spontaneous expressions of compound emotion in the wild. Later, these studies were used to define a new model of the production and perception of facial expressions (Martinez, 2017b).

**Exploratory design**

Although hypothesis-based has been the method of choice for decades, computational analyses now provide a mechanism to answer questions that were previously impossible to tackle. Some of the basic scientific questions I listed in the introduction of this paper, for instance, may not be properly addressed using a hypothesis-based approach. For example, a fundamental question in the study of the production and perception of facial expressions is to determine the number of expressions used across cultures (Martinez, 2017b; Barrett et al., in press). A hypothesis-based experimental design is unsuited to answer this question. We could define a study to test whether the six so-called “basic” emotions are used across cultures, but this would still not give us the actual number of cross-cultural expressions we want to know. An exploratory approach, however, does offer a method to properly address this. A very large database of images and videos of facial expressions collected in many cultures around the world, for instance, can be automatically analyzed to identify the facial configurations that are common across cultures. These results can then be used in behavioral experiments to test whether these facial configurations do indeed have a common interpretation across languages (Srinivasan & Martinez, 2019).

Exploratory experiments may be especially valuable for developmental studies, since these allow us to explore the evolution of our variables of interest over time. For example, we can use computer vision algorithms to delineate the narrowing of facial configurations as we age, or the acquisition of expertise on the production of expressions used in non-verbal communication.

Computing statistical significance is particularly important in exploratory experiments. Using the evaluation methods defined above does not mean we should not compute statistical significance. Given a large dataset of images or videos, a t-test can be readily computed, in which case, we should aim for p<.001, or use confident intervals (Cumming, 2013). Alternatively, we can compute the likelihood that the results we observed cannot be obtained from permutated data. To perform this test, we permute the labels of our AU detector and run the
same statistical analysis we used to complete our study. If we can still obtain results, regardless of whether these are the same or different than those obtained before permuting the AU labels, then our results should not be trusted, since any meaningless assignment of AUs yields a possible result.

**Which computer vision or machine learning approach?**

There is a good reason why, earlier, we defined the different approaches to automatic coding of AUs, because this will now facilitate the selection of our algorithm. In general, algorithms that use discriminant analysis (You et al., 2011) and deep learning (Benitez-Quiroz et al., 2017a, 2017b) work best for images. If there is large 3D head movement, algorithms that utilize structure from motion and 3D shape analysis may be preferred; or deep neural networks that can recover the 3D shape of the face from a single 2D image (Zhao et al., 2018). And, if we wish to identify facial movements of non-salient facial components, then 3D dense shape recovery methods would be preferred. In some cases, it may also be appropriate to select other algorithms. For example, if we wish to study the perception of implied motion in still images with AUs, then an algorithm that computes the optical flow or uses a template matching procedure to determine the movement of a set of fiducial points defining that AU may be the most appropriate.

If we are interested in video, the latest algorithm is that of Benitez-Quiroz et al. (2019), which has been shown to outperform other methods on standard datasets. But, as always, you will need to evaluate whether this (or some other algorithm) works best on your data.

We may also want to compare the results of an AU analysis with those of the changes in facial color due to emotional experiences. Facial articulations and color are believed to be controlled by (at least partially) dissociated neural mechanisms (Benitez-Quiroz et al., 2018), suggesting parallel ways of studying emotion. Having multiple means of investigating the same scientific question can add robustness to our studies.

Finally, if we select an algorithm, test it, and determine it does not work well on our data, we should then select an algorithm that uses a distinct approach. The reason for this is simple. Although multiple algorithms have been derived for each approach, if one of them does not work on the type of data we have, it is unlikely another algorithm in the same group will work much better. We have a better chance with an algorithm that uses a distinct methodology.
As detailed earlier in this paper, some of the computer vision and all the machine learning algorithms can be retrained with a portion of your data. That means you will need to manually annotate a portion of your still images and/or video sequences and then use them to retrain the available algorithm. This should only be done if none of the existing (pre-trained) algorithms worked and we have a computer vision expert that can help us perform this technical step, but it is an option to consider and one that typically yields excellent results.

Also note that some of these algorithms are available from companies, and may be easy to use, as out of the box tools. But others may be available from researchers and require basic computer science knowledge on how to operate them. As in the above, the best course of action is to add a computer vision researcher in your group that knows how to run and test these algorithms. But do not leave all the work to them. You should discuss which of the algorithms and approaches described in this paper are available to you and why one is believed to be a better choice than the others. Make sure you discuss how to evaluate the selected algorithms too.

**Conclusions**

Automatic facial action coding has the potential to be of major help to researchers studying the role faces play in a number of verbal and non-verbal social interactions (Martinez, 2017a; Benitez-Quiroz et al., 2014, 2016). This is especially useful for developmental psychologists interested in probing the role of facial configurations in a number of infant and developmental studies. Herein, I have summarized the main computer vision algorithms available to researchers and, most importantly, how to properly use them in scientific studies.

Several researchers are already using these algorithms in their research studies (Zanette et al., 2018; Martinez, 2017a; Sikka et al., 2015; De la Torre & Cohn, 2011), a number that is expected to grow rapidly in the next few years. While such systems are welcomed and a good opportunity to advance research in facial expressions, emotion, affect, sign language, and developmental psychology, they also need to be used and tested properly before being embraced as a universal solution to each and every facial analysis we might need. This paper provides a guide on how to achieve that. Specifically, I have presented a methodology to help researchers select the most appropriate computer vision algorithm for a given task and provided details of the distinct algorithms that are available to researchers. Taxonomies of the analyses and computer vision algorithms were presented in Figures 2 and 4.
It is also important to note that this paper provides a guide on how to use available facial action coding algorithms, not systems that purport to automatically detect emotion categories and valence. There is a good reason for this: the latter systems do not recognize all emotion categories or valence in images, but, instead, analyze images based on preconceived ideas of emotion that are most likely inaccurate (Barrett et al., in press). For example, Srinivasan & Martinez (2019) recently showed there are at least 17 facial expressions of happiness with a varying AUs, and these are not accounted for in computer algorithms designed to categorize emotion in images. The same is true for valence. Additionally, facial color is a marker of affect that has until very recently been omitted (Benitez-Quiroz et al., 2018) and omitting it can readily result in misinterpretations of the observed expressions.

In summary, the time is right to move to an automatic analysis of facial expressions. This is likely to revolutionize the study of nonverbal communication and emotion and will surely be a fundamental tool for developmental psychologist for years to come. However, when using these computational tools, care needs to be taken in both the selection of the computer vision algorithms and the experimental design, otherwise we run the risk of uncovering nonexistent features of our social and cognitive development.
References


Transactions on pattern analysis and machine intelligence. DOI: 10.1109/TPAMI.2018.2868952


Proceedings of the Royal Society of London B: Biological Sciences, 267(1459), 2239-2245. DOI: 10.1098/rspb.2000.1274


Martinez, A. M., & Kak, A. C. (2001). Pca versus lda. IEEE Transactions on Pattern Analysis & Machine Intelligence, (2), 228-233. DOI: 10.1109/34.908974


Reeb-Sutherland, B. C., Rankin Williams, L., Degnan, K. A., Pérez-Edgar, K., Chronis-Tuscano, A., Leibenluft, E., ... & Fox, N. A. (2015). Identification of emotional facial expressions


prosocial lies. Journal of Experimental Child Psychology, 150, 165-179. DOI: 10.1016/j.jecp.2016.05.007


Figure 1. a. Action Units (AUs). ©Dirk W. Eilert, Eilert-Academy, Germany. Reprinted with permission. b. Automatic annotation of AUs with the algorithm of Benitez-Quiroz et al. (2016). The individual whose face appears here gave signed consent for his likeness to be published in this article.
Figure 2. Can I use a computer vision system to automatically annotate action units in my images and videos? Shown here is a taxonomy of what computer vision algorithm can and cannot do at the moment. Follow the arrows to determine to which degree algorithms can help in your studies. If you reach the blue box, you will most likely be able to identify an algorithm that can do most (if not all) of the job (see text for how to identify the algorithm). If you reach the green box, there is likely an algorithm you can use, but you will need an expert AU coder to determine how well the algorithm works on your data and verify the results of your experiment. But, if you reach the red box, computer vision algorithms will provide only minimal help and you will require a human expert to aid, adjust and verify the algorithm and data at each stage of the experiment.
Figure 3. What coding is needed for our study? a. Qualitative coding of AUs with or without intensities. b. Quantitative analysis of AU activation over time. The individuals whose face appears here gave signed consent for their likeness to be published in this article.
**Figure 4.** A taxonomy of the most popular techniques used to automatically annotate action units in face images. Several algorithms use more than one of these techniques to detect AUs in face images.

**Figure 5.** The lines in the left image indicate the apparent movement of fiducial points (i.e., optical flow) needed to move from a norm (average) neutral face to the facial expression shown on right. Note how this optical flow specifies the outer pulling of the corners of the lips (AU 12) and the parting of the lips (AU 25). The individual whose face appears here gave signed consent for his likeness to be published in this article.
Figure 6. 2D and 3D shape automatically extracted from a video sequence using: a. Procrustes analysis with rotation invariant kernels, and b. non-rigid structure from motion. The individual whose face appears here gave signed consent for their likeness to be published in this article.

Figure 7. Given the image shown on left (indicated with a green frame), the algorithm of Pumerola et al (2018) is able to edit the image to illustrate what the face would look like with distinct AUs active at different intensities. Here, AUs 12 and 25 are added, with their intensities increasing from left to right. Note that the only real image is the one on left; all others are computer generated, i.e., fake. Adapted with permission from Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In Proceedings of the European Conference on Computer Vision (pp. 818-833).