Journal of Machine Learning Research 13 (2012) 1589-1608 Submitted 10/11; Revised 2/12; Published 5/12

A Model of the Perception of Facial Expressions of Emotion by Humans: Research Overview and Perspectives

Aleix Martinez ALEIX@ECE.OSU.EDU
Shichuan Du DUS@ECE.OSU.EDU
Department of Electrical and Computer Engineering
The Ohio State University
2015 Neil Avenue
Columbus, OH 43210, USA

Editors: Isabelle Guyon and Vassilis Athitsos

Abstract

In cognitive science and neuroscience, there have been two leading models describing how humans perceive and classify facial expressions of emotion—the continuous and the categorical model. The continuous model defines each facial expression of emotion as a feature vector in a face space. This model explains, for example, how expressions of emotion can be seen at different intensities. In contrast, the categorical model consists of C classifiers, each tuned to a specific emotion category. This model explains, among other findings, why the images in a morphing sequence between a happy and a surprise face are perceived as either happy or surprise but not something in between. While the continuous model has a more difficult time justifying this latter finding, the categorical model is not as good when it comes to explaining how expressions are recognized at different intensities or modes. Most importantly, both models have problems explaining how one can recognize combinations of emotion categories such as happily surprised versus angrily surprised versus surprise. To resolve these issues, in the past several years, we have worked on a revised model that justifies the results reported in the cognitive science and neuroscience literature. This model consists of C distinct continuous spaces. Multiple (compound) emotion categories can be recognized by linearly combining these C face spaces. The dimensions of these spaces are shown to be mostly configural. According to this model, the major task for the classification of facial expressions of emotion is precise, detailed detection of facial landmarks rather than recognition. We provide an overview of the literature justifying the model, show how the resulting model can be employed to build algorithms for the recognition of facial expression of emotion, and propose research directions in machine learning and computer vision researchers to keep pushing the state of the art in these areas. We also discuss how the model can aid in studies of human perception, social interactions and disorders.

Keywords: vision, face perception, emotions, computational modeling, categorical perception, face detection

1. Introduction

The face is an object of major importance in our daily lives. Faces tell us the identity of the person we are looking at and provide information on gender, attractiveness and age, among many others. Of primary interest is the production and recognition of facial expressions of emotion. Emotions play a fundamental role in human cognition (Damasio, 1995) and are thus essential in studies of cognitive science, neuroscience and social psychology. Facial expressions of emotion could also

play a pivotal role in human communication (Schmidt and Cohn, 2001). And, sign languages use facial expressions to encode part of the grammar (Wilbur, 2011). It has also been speculated that expressions of emotion were relevant in human evolution (Darwin, 1872). Models of the perception of facial expressions of emotion are thus important for the advance of many scientific disciplines.

A first reason machine learning and computer vision researchers are interested in creating computational models of the perception of facial expressions of emotion is to aid studies in the above sciences (Martinez, 2003). Furthermore, computational models of facial expressions of emotion are important for the development of artificial intelligence (Minsky, 1988) and are essential in human-computer interaction (HCI) systems (Pentland, 2000).

Yet, as much as we understand how facial expressions of emotion are produced, very little is known on how they are interpreted by the human visual system. Without proper models, the scientific studies summarized above as well as the design of intelligent agents and efficient HCI platforms will continue to elude us. A HCI system that can easily recognize expressions of no interest to the human user is of limited interest. A system that fails to recognize emotions readily identified by us is worse.

In the last several years, we have defined a computational model consistent with the cognitive science and neuroscience literature. The present paper presents an overview of this research and a perspective of future areas of interest. We also discuss how machine learning and computer vision should proceed to successfully emulate this capacity in computers and how these models can aid in studies of visual perception, social interactions and disorders such as schizophrenia and autism. In particular, we provide the following discussion.

- A model of human perception of facial expressions of emotion: We provide an overview of the cognitive science literature and define a computational model consistent with it.

- Dimensions of the computational space: Recent research has shown that human used mostly shape for the perception and recognition of facial expressions of emotion. In particular, we show that configural features are of much use in this process. A configural feature is defined as a non-rotation invariant modeling of the distance between facial components; for example, the vertical distance between eyebrows and mouth.

- We argue that to overcome the current problems of face recognition algorithms (including identity and expressions), the area should make a shift toward a more shape-based modeling. Under this model, the major difficulty for the design of computer vision and machine learning systems is that of precise detection of the features, rather than classification. We provide a perspective on how to address these problems.

The rest of the paper is organized as follows. Section 2 reviews relevant research on the perception of facial expressions of emotion by humans. Section 3 defines a computational model consistent with the results reported in the previous section. Section 4 illustrates the importance of configural and shape features for the recognition of emotions in face images. Section 5 argues that the real problem in machine learning and computer vision is a detection one and emphasizes the importance of research in this domain before we can move forward with improved algorithms of face recognition. In Section 6, we summarize some of the implications of the proposed model. We conclude in Section 7.

2. Facial Expressions: From Production to Perception

The human face is an engineering marvel. Underneath our skin, a large number of muscles allow us to produce many configurations. The face muscles can be summarized as Action Unit (AU) (Ekman and Friesen, 1976) defining positions characteristic of facial expressions of emotion. These face muscles are connected to the motor neurons in the cerebral cortex through the corticobulbar track. The top muscles are connected bilaterally, while the bottom ones are connected unilaterally to the opposite hemisphere. With proper training, one can learn to move most of the face muscles independently. Otherwise, facial expressions take on predetermined configurations.

There is debate on whether these predetermined configurations are innate or learned (nature vs. nurture) and whether the expressions of some emotions is universal (Izard, 2009). By universal, we mean that people from different cultures produce similar muscle movements when expressing some emotions. Facial expressions typically classified as universal are joy, surprise, anger, sadness, disgust and fear (Darwin, 1872; Ekman and Friesen, 1976). Universality of emotions is controversial, since it assumes facial expressions of emotion are innate (rather than culturally bound). It also favors a categorical perception of facial expressions of emotion. That is, there is a finite set of predefined classes such as the six listed above. This is known as the *categorical model*.

In the categorical model, we have a set of C classifiers. Each classifier is specifically designed to recognize a single emotion label, such as surprise. Several psychophysical experiments suggest the perception of emotions by humans is categorical (Ekman and Rosenberg, 2005). Studies in neuroscience further suggest that distinct regions (or pathways) in the brain are used to recognize different expressions of emotion (Calder et al., 2001).

An alternative to the categorical model is the *continuous model* (Russell, 2003; Rolls, 1990). Here, each emotion is represented as a feature vector in a multidimensional space given by some characteristics common to all emotions. One such model is Russell's 2-dimensional circumplex model (Russell, 1980), where the first basis measures pleasure-displeasure and the second arousal. This model can justify the perception of many expressions, whereas the categorical model needs to define a class (i.e., classifier) for every possible expression. It also allows for intensity in the perception of the emotion label. Whereas the categorical model would need to add an additional computation to achieve this goal (Martinez, 2003), in the continuous model the intensity is intrinsically defined in its representation. Yet, morphs between expressions of emotions are generally classified to the closest class rather than to an intermediate category (Beale and Keil, 1995). Perhaps more interestingly, the continuous model better explains the caricature effect (Rhodes et al., 1987; Calder et al., 1997), where the shape features of someone's face are exaggerated (e..g, making a long nose longer). This is because the farther the feature vector representing that expression is from the mean (or center of the face space), the easier it is to recognize it (Valentine, 1991).

In neuroscience, the multidimensional (or continuous) view of emotions was best exploited under the limbic hypothesis (Calder et al., 2001). Under this model, there should be a neural mechanism responsible for the recognition of all facial expressions of emotion, which was assumed to take place in the limbic system. Recent results have however uncovered dissociated networks for the recognition of most emotions. This is not necessarily proof of a categorical model, but it strongly suggests that there are at least distinct groups of emotions, each following distinct interpretations.

Furthermore, humans are only very good at recognizing a number of facial expressions of emotion. The most readily recognized emotions are happiness and surprise. It has been shown that joy and surprise can be robustly identified extremely accurately at almost any resolution (Du and Mar-

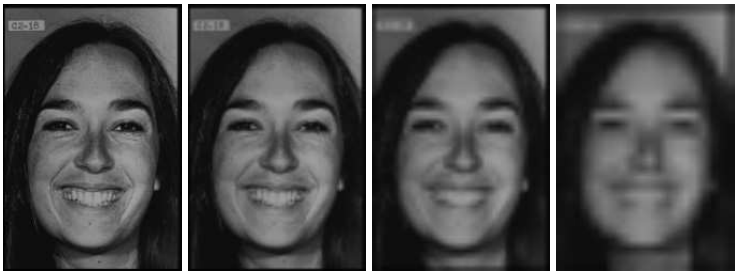


Figure 1: Happy faces at four different resolutions. From left o right: 240 by 160, 120 by 80, 60 by 40, and 30 by 20 pixels. All images have been resized to a common image size for visualization.

tinez, 2011). Figure 1 shows a happy expression at four different resolutions. The reader should not have any problem recognizing the emotion in display even at the lowest of resolutions. However, humans are not as good at recognizing anger and sadness and are even worse at fear and disgust.

A major question of interest is the following. Why are some facial configurations more easily recognizable than others? One possibility is that expressions such as joy and surprise involve larger face transformations than the others. This has recently proven not to be the case (Du and Martinez, 2011). While surprise does have the largest deformation, this is followed by disgust and fear (which are poorly recognized). Learning why some expressions are so readily classified by our visual system should facilitate the definition of the form and dimensions of the computational model of facial expressions of emotion.

The search is on to resolve these two problems. First, we need to determine the *form* of the computational space (e.g., a continuous model defined by a multidimensional space). Second, we ought to define the *dimensions* of this model (e.g., the dimensions of this multidimensional face space are given by configural features). In the following sections we overview the research we have conducted in the last several years leading to a solution to the above questions. We then discuss on the implications of this model. In particular, we provide a perspective on how machine learning and computer vision researcher should move forward if they are to define models based on the perception of facial expressions of emotion by humans.

3. A Model of the Perception of Facial Expressions of Emotion

In cognitive science and neuroscience researchers have been mostly concerned with models of the perception and classification of the six facial expressions of emotion listed above. Similarly, computer vision and machine learning algorithms generally employ a face space to represent these six emotions. Sample feature vectors or regions of this feature space are used to represent each of these six emotion labels. This approach has a major drawback—it can only detect one emotion from a single image. In machine learning, this is generally done by a winner-takes-all approach (Torre and Cohn, 2011). This means that when a new category wants to be included, one generally needs to provide labeled samples of it to the learning algorithm.

Yet, everyday experience demonstrates that we can perceive more than one emotional category in a single image (Martinez, 2011), even if we have no prior experience with it. For example,

Figure 2: Faces expressing different surprise. From left to right: happily surprised, sadly surprised, angrily surprised, fearfully surprised, disgustedly surprised, and surprise.

Figure 2 shows images of faces expressing different surprises—happily surprised, angrily surprised, fearfully surprised, disgustedly surprised and the typically studied surprise.

If we were to use a continuous model, we would need to have a very large number of labels represented all over the space; including all possible types of surprises. This would require a very large training set, since each possible combination of labels would have to be learned. But this is the same problem a categorical model would face. In such a case, dozens if not hundreds of sample images for each possible category would be needed. Alternatively, Susskind et al. (2007) have shown that the appearance of a continuous model may be obtained from a set of classifiers defining a small number of categories.

If we define an independent computational (face) space for a small number of emotion labels, we will only need sample faces of those few facial expressions of emotion. This is indeed the approach we have taken. Details of this model are given next.

Key to this model is to note that we can define new categories as linear combinations of a small set of categories. Figure 3 illustrates this approach. In this figure, we show how we can obtain the above listed different surprises as a linear combination of known categories. For instance, happily surprised can be defined as expressing 40% joy plus 60% surprise, that is, expression = .4 happy + .6 surprise. A large number of such expressions exist that are a combination of the six emotion categories listed above and, hence, the above list of six categories is a potential set of basic emotion classes. Also, there is some evidence form cognitive science to suggest that these are important categories for humans (Izard, 2009) Of course, one needs not base the model on this set of six emotions. This is an area that will undoubtedly attract lots of interest. A question of particular interest is to determine not only which basic categories to include in the model but how many. To this end both, cognitive studies with humans and computational extensions of the proposed model will be necessary, with the results of one area aiding the research of the other.

The approach described in the preceding paragraph would correspond to a categorical model. However, we now go one step further and define each of these face spaces as continuous feature spaces, Figure 3. This allows for the perception of each emotion at different intensities, for example, less happy to exhilarant (Neth and Martinez, 2010). Less happy would correspond to a feature vector (in the left most face space in the figure) closer to the mean (or origin of the feature space). Feature vectors farther from the mean would be perceived as happier. The proposed model also explains the caricature effect, because within each category the face space is continuous and exaggerating the

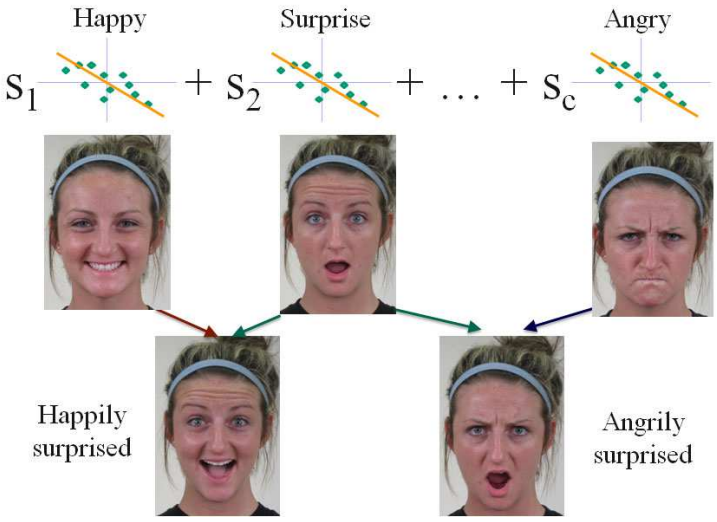


Figure 3: This figure shows how to construct linear combinations of known categories. At the top of the figure, we have the known or learned categories (emotions). The coefficients s_i determine the contribution of each of these categories to the final perception of the emotion.

expression will move the feature vector representing the expression further from the mean of that category.

Furthermore, the proposed model can define new terms, for example, "hatred" which is defined as having a small percentage of disgust and a larger percentage of anger; still linear. In essence, the intensity observed in this *continuous representation* defines the weight of the contribution of each basic category toward the final decision (classification). It also allows for the representation and recognition of a very large number of emotion categories without the need to have a categorical space for each or having to use many samples of each expression as in the continuous model.

The proposed model thus bridges the gap between the categorical and continuous ones and resolves most of the debate facing each of the models individually. To complete the definition of the model, we need to specify what defines each of the dimensions of the continuous spaces representing each category. We turn to this problem in the next section.

4. Dimensions of the Model

In the early years of computer vision, researchers derived several feature- and shape-based algorithms for the recognition of objects and faces (Kanade, 1973; Marr, 1976; Lowe, 1983). In these methods, geometric, shape features and edges were extracted from an image and used to build a model of the face. This model was then fitted to the image. Good fits determined the class and position of the face.

Later, the so-called appearance-based approach, where faces are represented by their pixel-intensity maps or the response of some filters (e.g., Gabors), was studied (Sirovich and Kirby,

1987). In this alternative texture-based approach, a metric is defined to detect and recognize faces in test images (Turk and Pentland, 1991). Advances in pattern recognition and machine learning have made this the preferred approach in the last two decades (Brunelli and Poggio, 1993).

Inspired by this success, many algorithms developed in computer vision for the recognition of expressions of emotion have also used the appearance-based model (Torre and Cohn, 2011). The appearance-based approach has also gained momentum in the analysis of AUs from images of faces. The main advantage of the appearance-based model is that one does not need to predefine a feature or shape model as in the earlier approaches. Rather, the face model is inherently given by the training images.

The appearance-based approach does provide good results from near-frontal images of a reasonable quality, but it suffers from several major inherent problems. The main drawback is its sensitivity to image manipulation. Image size (scale), illumination changes and pose are all examples of this. Most of these problems are intrinsic to the definition of the approach since this cannot generalize well to conditions not included in the training set. One solution would be to enlarge the number of training images (Martinez, 2002). However, learning from very large data sets (in the order of millions of samples) is, for the most part, unsolved (Lawrence, 2005). Progress has been made in learning complex, non-linear decision boundaries, but most algorithms are unable to accommodate large amounts of data—either in space (memory) or time (computation).

This begs the question as to how the human visual system solves the problem. One could argue that, throughout evolution, the homo genus (and potentially before it) has been exposed to trillions of faces. This has facilitated the development of simple, yet robust algorithms. In computer vision and machine learning, we wish to define algorithms that take a shorter time to learn a similarly useful image representation. One option is to decipher the algorithm used by our visual system. Research in face recognition of identity suggests that the algorithm used by the human brain is not appearance-based (Wilbraham et al., 2008). Rather, it seems that, over time, the algorithm has identified a set of robust features that facilitate rapid categorization (Young et al., 1987; Hosie et al., 1988; Barlett and Searcy, 1993).

This is also the case in the recognition of facial expressions of emotion (Neth and Martinez, 2010). Figure 4 shows four examples. These images all bear a neutral expression, that is, an expression associated to no emotion category. Yet, human subjects perceive them as expressing sadness, anger, surprise and disgust. The most striking part of this illusion is that these faces do not and cannot express any emotion, since all relevant AUs are inactive. This effect is called over-generalization (Zebrowitz et al., 2010), since human perception is generalizing the learned features defining these face spaces over to images with a different label.

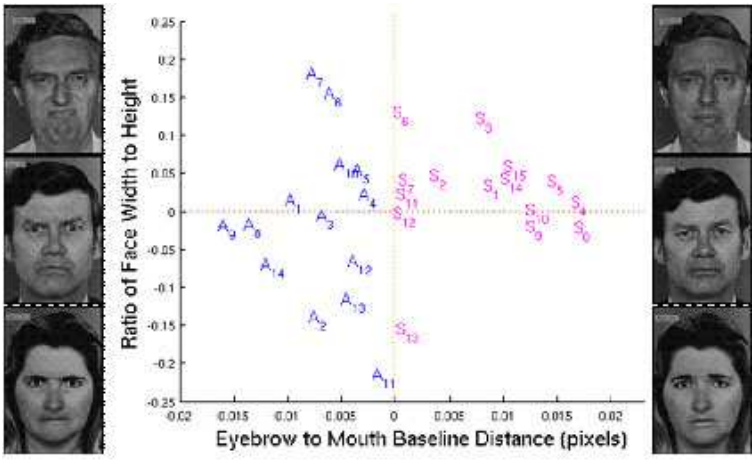
The images in Figure 4 do have something in common though—they all include a configural transformation. What the human visual system has learned is that faces do not usually look like those in the image. Rather the relationship (distances) between brows, nose, mouth and the contour of the face is quite standard. They follow a Gaussian distribution with small variance (Neth and Martinez, 2010). The images shown in this figure however bear uncanny distributions of the face components. In the sad-looking example, the distance between the brows and mouth is larger than normal (Neth and Martinez, 2009) and the face is thinner than usual (Neth and Martinez, 2010). This places this sample face, most likely, outside the 99% confidence interval of all Caucasian faces on these two measures. The angry-looking face has a much-shorter-than-average brow to mouth distance and a wide face. While the surprise-looking face has a large distance between eyes and

Figure 4: The four face images and schematics shown above all correspond to neutral expressions (i.e., the sender does not intend to convey any emotion to the receiver). Yet, most human subjects interpret these faces as conveying anger, sadness, surprise and disgust. Note that although these faces look very different from one another, three of them are actually morphs from the same (original) image.

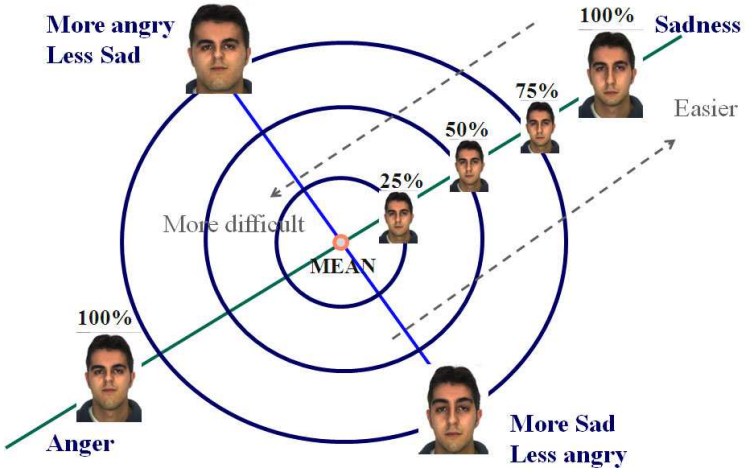
brows and a thinner face. The disgust-looking face has a shorter distance between brows, eyes, nose and mouth. These effects are also clear in the schematic faces shown in the figure.

Yet, configural cues alone are not sufficient to create an impressive, lasting effect. Other shape changes are needed. For example, the curvature of the mouth in joy or the opening of the eyes—showing additional sclera—in surprise. Note how the surprise-looking face in Figure 4 appears to also express disinterest or sleepiness. Wide-open eyes would remove these perceptions. But this can only be achieved with a shape change. Hence, our face spaces should include both, configural and shape features. It is important to note that configural features can be obtained from an appropriate representation of shape. Expressions such as fear and disgust seem to be mostly (if not solely) based on shape features, making recognition less accurate and more susceptible to image manipulation. We have previously shown (Neth and Martinez, 2010) that configural cues are amongst the most discriminant features in a classical (Procrustes) shape representation, which can be made invariant to 3D rotations of the face (Hamsici and Martinez, 2009a).

Thus, each of the six categories of emotion (happy, sad, surprise, angry, fear and disgust) is represented in a shape space given by classical statistical shape analysis. First the face and the shape of the major facial components are automatically detected. This includes delineating the brows, eyes, nose, mouth and jaw line. The shape is then sample with d equally spaced landmark points. The mean (center of mass) of all the points is computed. The $2d$-dimensional shape feature vector is given by the x and y coordinates of the d shape landmarks subtracted by the mean and divided by its norm. This provides invariance to translation and scale. 3D rotation invariance can be achieved with the inclusion of a kernel as defined in Hamsici and Martinez (2009a). The dimensions of each emotion category can now be obtained with the use of an appropriate discriminant analysis method. We use the algorithm defined by Hamsici and Martinez (2008) because it minimizes the Bayes classification error.



(a)



(b)

Figure 5: (a) Shown here are the two most discriminant dimensions of the face shape vectors. We also plot the images of anger and sadness of Ekman and Friesen (1976). In dashed are simple linear boundaries separating angry and sad faces according to the model. The first dimension (distance between brows and mouth) successfully classifies 100% of the sample images. This continuous model is further illustrated in (b). Note that, in the proposed computational model, the face space defining sadness corresponds to the right-bottom quadrant, while that of anger is given by the left-top quadrant. The dashed arrows in the figure reflect the fact that as we move away from the "mean" (or norm) face, recognition of that emotion become easier.

As an example, the approach detailed in this section identifies the distance between the brows and mouth and the width of the face as the two most important shape features of anger and sadness. It is important to note that, if we reduce the computational spaces of anger and sadness to 2-dimensions, they are almost indistinguishable. Thus, it is possible that these two categories are in fact connected by a more general one. This goes back to our question of the number of basic categories used by the human visual system. The face space of anger and sadness is illustrated in Figure 5, where we have also plotted the feature vectors of the face set of Ekman and Friesen (1976).

As in the above, we can use the shape space defined above to find the two most discriminant dimensions separating each of the six categories listed earlier. The resulting face spaces are shown in Figure 6. In each space, a simple linear classifier in these spaces can successfully classify each emotion very accurately. To test this, we trained a linear support vector machine (Vapnik, 1998) and use the leave-one-out test on the data set of images of Ekman and Friesen (1976). Happiness is correctly classified 99% of the time. Surprise and disgust 95% of the time. Sadness 90% and anger 94%. While fear is successfully classified at 92%. Of course, adding additional dimensions in the feature space and using nonlinear classifiers can readily achieve perfect classification (i.e., 100%). The important point from these results is to note that simple configural features can *linearly* discriminate most of the samples in each emotion. These features are very robust to image degradation and are thus ideal for recognition in challenging environments (e.g., low resolution)—a message to keep in mind for the development of machine learning and computer vision systems.

5. Precise Detection of Faces and Facial Features

As seen thus far, human perception is extremely tuned to small configural and shape changes. If we are to develop computer vision and machine learning systems that can emulate this capacity, the real problem to be addressed by the community is that of *precise detection of faces and facial features* (Ding and Martinez, 2010). Classification is less important, since this is embedded in the detection process; that is, we want to precisely detect changes that are important to recognize emotions.

Most computer vision algorithms defined to date provide, however, inaccurate detections. One classical approach to detection is template matching. In this approach, we first define a template (e.g., the face or the right eye or the left corner of the mouth or any other feature we wish to detect). This template is learned from a set of sample images; for example, estimating the distribution or manifold defining the appearance (pixel map) of the object (Yang et al., 2002). Detection of the object is based on a window search. That is, the learned template is compared to all possible windows in the image. If the template and the window are similar according to some metric, then the bounding box defining this window marks the location and size (scale) of the face. The major drawback of this approach is that it yields imprecise detections of the learned object, because a window of an non-centered face is more similar to the learned template than a window with background (say, a tree). An example of this result is shown in Figure 7.

A solution to the above problem is to learn to discriminate between non-centered windows of the objects and well centered ones (Ding and Martinez, 2010). In this alternative, a non-linear classifier (or some density estimator) is employed to discriminate the region of the feature space defining well-centered windows of the objects and non-centered ones. We call these non-centered windows the context of the object, in the sense that these windows provide the information typically found around the object but do not correspond to the actual face. This features versus context idea is illustrated in Figure 8. This approach can be used to precisely detect faces, eyes, mouth, or any

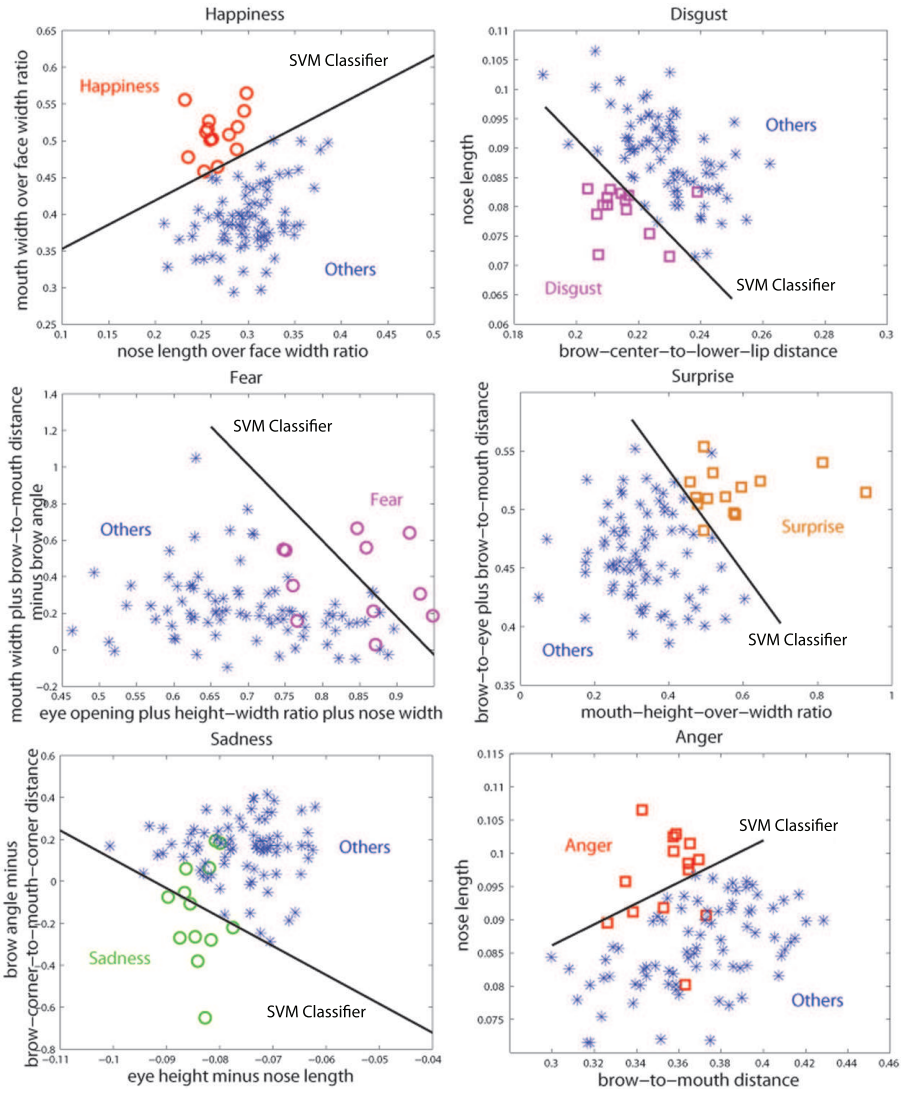


Figure 6: Shown in the above are the six feature spaces defining each of the six basic emotion categories. A simple linear Support Vector Machine (SVM) can achieve high classification accuracies; where we have used a one-versus-all strategy to construct each classifier and tested it using the leave-one-out strategy. Here, we only used two features (dimensions) for clarity of presentation. Higher accuracies are obtained if we include additional dimensions and training samples.

Figure 7: Two example of imprecise detections of a face with a state of the art algorithm.

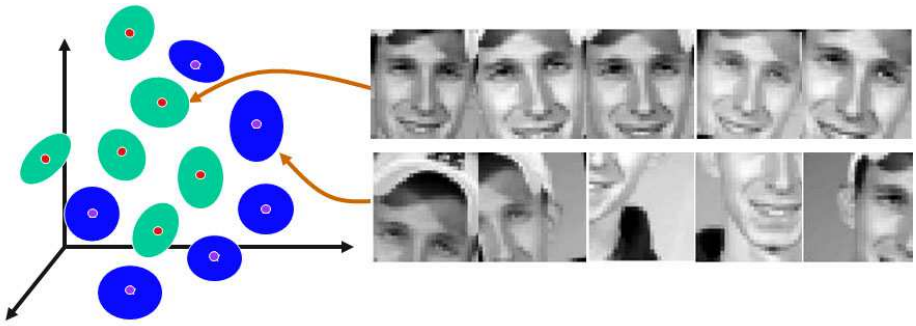


Figure 8: The idea behind the features versus context approach is to learn to discriminate between the feature we wish to detect (e.g., a face, an eye, etc.) and poorly detected versions of it. This approach eliminates the classical overlapping of multiple detections around the object of interest at multiple scales. At the same time, it increases the accuracy of the detection because we are moving away from poor detections and toward precise ones.

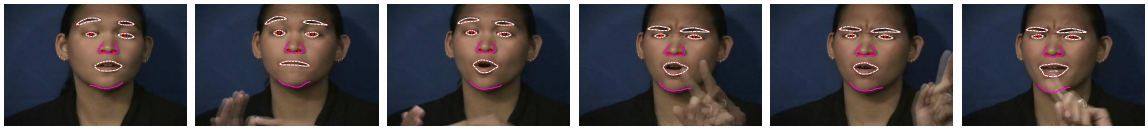


Figure 9: Precise detections of faces and facial features using the algorithm of (Ding and Martinez, 2010).

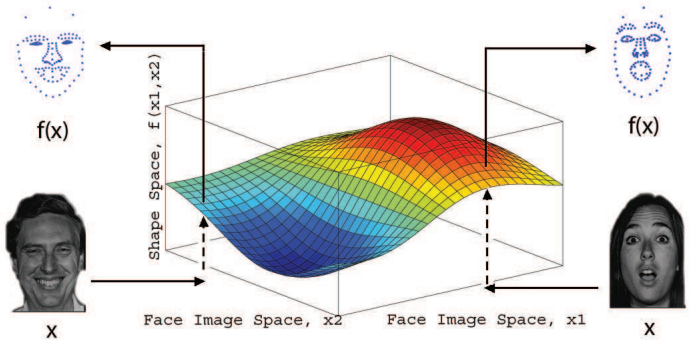


Figure 10: Manifold learning is ideal for learning mappings between face (object) images and their shape description vectors.

other facial feature where there is a textural discrimination between it and its surroundings. Figure 9 shows some sample results of accurate detection of faces and facial features with this approach.

The same features versus context idea can be applied to other detection and modeling algorithms, such as Active Appearance Models (AAM) (Cootes et al., 2001). AAM use a linear model—usually based on Principal Component Analysis (PCA)—to learn the relationship between the shape of an object (e.g., a face) and its texture. One obvious limitation is that the learned model is linear. A solution to this problem is to employ a kernel map. Kernel PCA is one option. Once we have introduced a kernel we can move one step further and use it to address additional issues of interest. A first capability we may like to add to a AAM is the possibility to work with three-dimensions. The second could be to omit the least-squares iterative nature of the Procrustes alignment required in most statistical shape analysis methods such as AAM. An approach that successfully addresses these problem uses a set of kernels called Rotation Invariant Kernels (RIK) (Hamsici and Martinez, 2009a). RIK add yet another important advantage to shape analysis: they provide rotation invariance. Thus, once the shape is been mapped to the RIK space, objects (e.g., faces) are invariant to translation, scale and rotation. These kernels are thus very attractive for the design of AAM algorithms (Hamsici and Martinez, 2009b).

By now we know that humans are very sensitive to small changes. But we do not yet know how sensitive (or accurate). Of course, it is impossible to be pixel accurate when marking the boundaries of each facial feature, because edges blur over several pixels. This can be readily observed by zooming in the corner of an eye. To estimate the accuracy of human subjects, we performed the following experiment. First, we designed a system that allows users to zoom in at any specified location to facilitate delineation of each of the facial features manually. Second, we asked three people (herein referred to as judges) to manually delineate each of the facial components of close to $4,000$ images of faces. Third, we compared the markings of each of the three judges. The within-judge variability was (on average) 3.8 pixels, corresponding to a percentage of error of 1.2% in terms of the size of the face. This gives us an estimate of the accuracy of the manual detections. The average error of the algorithm of Ding and Martinez (2010) is 7.3 pixels (or 2.3%), very accurate but still far short of what humans can achieve. Thus, further research is needed to develop computer vision algorithms that can extract even more accurate detection of faces and its components.

Figure 11: Shape detection examples at different resolutions. Note how the shape estimation is almost as good regardless of the resolution of the image.

Another problem is what happens when the resolution of the image diminishes. Humans are quite robust to these image manipulations (Du and Martinez, 2011). One solution to this problem is to use manifold learning. In particular, we wish to define a non-linear mapping $f(.)$ between the image of a face and its shape. This is illustrated in Figure 10. That is, given enough sample images and their shape feature vectors described in the preceding section, we need to find the function which relates the two. This can be done, for example, using kernel regression methods (Rivera and Martinez, 2012). One of the advantages of this approach is that this function can be defined to detect shape from very low resolution images or even under occlusions. Occlusions can be "learned" by adding synthetic occlusions or missing data in the training samples but leaving the shape feature vector undisturbed (Martinez, 2002). Example detections using this approach are shown in Figure 11.

One can go one step further and recover the three-dimensional information when a video sequence is available (Gotardo and Martinez, 2011a). Recent advances in non-rigid structure from motion allow us to recover very accurate reconstructions of both the shape and the motion even under occlusion. A recent approach resolves the nonlinearity of the problem using kernel mappings (Gotardo and Martinez, 2011b).

Combining the two approaches to detection defined in this section should yield even more accurate results in low-resolution images and under occlusions or other image manipulations. We hope that more research will be devoted to this important topic in face recognition.

The approaches defined in this section are a good start, but much research is needed to make these systems comparable to human accuracies. We argue that research in machine learning should address these problems rather than the typical classification one. A first goal is to define algorithms

that can detect face landmarks very accurately even at low resolutions. Kernel methods and regression approaches are surely good solutions as illustrated above. But more targeted approaches are needed to define truly successful computational models of the perception of facial expressions of emotion.

6. Discussion

In the real world, occlusions and unavoidable imprecise detections of the fiducial points, among others, are known to affect recognition (Torre and Cohn, 2011; Martinez, 2003). Additionally, some expressions are, by definition, ambiguous. Most importantly though seems to be the fact that people are not very good at recognizing facial expressions of emotion even under favorable condition (Du and Martinez, 2011). Humans are very robust at detection joy and surprise from images of faces; regardless of the image conditions or resolution. However, we are not as good at recognizing anger and sadness and are worst at fear and disgust.

The above results suggest that there could be three groups of expressions of emotion. The first group is intended for conveying emotions to observers. These expressions have evolved a facial construct (i.e., facial muscle positions) that is distinctive and readily detected by an observer at short or large distances. Example expressions in this group are happiness and surprise. A computer vision system—especially a HCI—should make sure these expressions are accurately and robustly recognized across image degradation. Therefore, we believe that work needs to be dedicated to make systems very robust when recognizing these emotions.

The second group of expressions (e.g., anger and sadness) is reasonably recognized at close proximity only. A computer vision system should recognize these expressions in good quality images, but can be expected to fail as the image degrades due to resolution or other image manipulations. An interesting open question is to determine why this is the case and what can be learned about human cognition from such a result.

The third and final group of emotions constitutes those at which humans are not very good recognizers. This includes expressions such as fear and disgust. Early work (especially in evolutionary psychology) had assumed that recognition of fear was primal because it served as a necessary survival mechanism (LeDoux, 2000). Recent studies have demonstrated much the contrary. Fear is generally poorly recognized by healthy human subjects (Smith and Schyns, 2009; Du and Martinez, 2011). One hypothesis is that expressions in this group have evolved for other than communication reasons. For example, it has been proposed that fear opens sensory channels (i.e., breathing in and wide open eye), while disgust closes them (i.e., breathing out and closed eyes) (Susskind et al., 2008). Under this model, the receiver has learned to identify those face configurations to some extent, but without the involvement of the sender—modifying the expression to maximize transmission of information through a noisy environment—the recognition of these emotions has remained poor. Note that people can be trained to detect such changes quite reliably (Ekman and Rosenberg, 2005), but this is not the case for the general population.

Another area that will require additional research is to exploit other types of facial expressions. Facial expressions are regularly used by people in a variety of setting. More research is needed to understand these. Moreover, it will be important to test the model in natural occurring environments. Collection and handling of this data poses several challenges, but the research described in these pages serves as a good starting point for such studies. In such cases, it may be necessary to go beyond a linear combination of basic categories. However, without empirical proof for the need

of something more complex than linear combinations of basic emotion categories, such extensions are unlikely. The cognitive system has generally evolved the simplest possible algorithms for the analysis or processing of data. Strong evidence of more complex models would need to be collected to justify such extensions. One way to do this is by finding examples that cannot be parsed by the current model, suggesting a more complex structure is needed.

It is important to note that these results will have many applications in studies of agnosias and disorders. Of particular interest are studies of depression or anxiety disorders. Depression afflicts a large number of people in the developed countries. Models that can help us better understand its cognitive processes, behaviors and patterns could be of great importance for the design of coping mechanisms. Improvements may also be possible if it were to better understand how facial expressions of emotion affect these people. Other syndromes such as autism are also of great importance these days. More children than ever are being diagnosed with the disorder (CDC, 2012; Prior, 2003). We know that autistic children do not perceive facial expressions of emotion as others do (Jemel et al., 2006) (but see Castelli, 2005). A modified computational model of the perception of facial expressions of emotion in autism could help design better teaching tools for this group and may bring us closer to understanding the syndrome.

There are indeed many great possibilities for machine learning researchers to help move these studies forward. Extending or modifying the modeled summarized in the present paper is one way. Developing machine learning algorithms to detect face landmark more accurately is another. Developing statistical tools that more accurately represent the underlying manifold or distribution of the data is yet another great way to move the state of the art forward.

7. Conclusions

In the present work we have summarized the development of a model of the perception of facial expressions of emotion by humans. A key idea in this model is to linearly combine a set of face spaces defining some basic emotion categories. The model is consistent with our current understanding of human perception and can be successfully exploited to achieve great recognition results for computer vision and HCI applications. We have shown how, to be consistent with the literature, the dimensions of these computational spaces need to encode configural and shape features.

We conclude that to move the state of the art forward, face recognition research has to focus on a topic that has received little attention in recent years—precise, detailed detection of faces and facial features. Although we have focused our study on the recognition of facial expressions of emotion, we believe that the results apply to most face recognition tasks. We have listed a variety of ways in which the machine learning community can get involved in this research project and briefly discussed applications in the study of human perception and the better understanding of disorders.

Acknowledgments

This research was supported in part by the National Institutes of Health, grants R01 EY 020834 and R21 DC 011081.

References

J. C. Barlett and J. Searcy. Inversion and configuration of faces. *Cognitive Psychology*, 25(3): 281–316, 1993.

J. M. Beale and F. C. Keil. Categorical effects in the perception of faces. *Cognition*, 57:217–239, 1995.

R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

A. J. Calder, A. W. Young, D. Rowland, and D. I. Perrett. Computer-enhanced emotion in facial expressions. *Proceedings of the Royal Society of London B*, 264:919–925, 1997.

A. J. Calder, A. D. Lawrence, and A. W. Young. Neuropsychology of fear and loathing. *Nature Review Neuroscience*, 2:352–363, 2001.

F. Castelli. Understanding emotions from standardized facial expressions in autism and normal development. *Autism*, 9:428–449, 2005.

CDC. Center for Disease Control and Prevention. Prevalence of autism spectrum disorders autism and developmental disabilities monitoring network, 14 sites, united states, 2008. *Morbidity and Mortality Weekly Report (MMWR)*, 61, 2012.

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

A. R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. G. P. Putnam's Sons, New York, 1995.

C. Darwin. *The Expression of the Emotions in Man and Animal*. J. Murray., London, 1872.

L. Ding and A. M. Martinez. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32:2022–2038, 2010.

S. Du and A. M. Martinez. The resolution of facial expressions of emotion. *Journal of Vision*, 11(13):24, 2011.

P. Ekman and W.V. Friesen. *Pictures of Facial Affect*. Consulting Psychologists Press, Palo Alto, CA, 1976.

P. Ekman and E.L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, New York, 2nd edition, 2005.

P. F. U. Gotardo and A. M. Martinez. Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2051–2065, 2011a.

P. F. U. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011b.

O. C. Hamsici and A. M. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30:647–657, 2008.

O. C. Hamsici and A. M. Martinez. Rotation invariant kernels and their application to shape analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31:1985–1999, 2009a.

O. C. Hamsici and A. M. Martinez. Active appearance models with rotation invariant kernels. In *IEEE Proc. International Conference on Computer Vision*, 2009b.

J. A. Hosie, H. D. Ellis, and N. D. Haig. The effect of feature displacement on the perception of well-known faces. *Perception*, 17(4):461–474, 1988.

C. E. Izard. Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual Review of Psychology*, 60:1–25, 2009.

B. Jemel, L. Mottron, and M. Dawson. Impaired face processing in autism: Fact or artifact? *Journal of Autism and Developmental Disorders*, 36:91–106, 2006.

T. Kanade. *Picture Processing System by Computer Complex and Recognition of Human Faces*. PhD thesis, Kyoto University, Japan, 1973.

N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, (6):1783–1816, 2005.

J.E. LeDoux. Emotion circuits in the brain. *Annual Review of Nueroscience*, 23:155–184, 2000.

D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1983.

D. Marr. Early processing of visual information. *Philosophical Transactions of the Royal Society of London*, 275(942):483–519, 1976.

A. M. Martinez. Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(6): 748–763, 2002.

A. M. Martinez. Matching expression variant faces. *Vision Research*, 43:1047–1060, 2003.

A. M. Martinez. Deciphering the face. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition, workshop*, 2011.

M. Minsky. *The Society of Mind*. Simon & Schuster, New York, N.Y., 1988.

D. Neth and A. M. Martinez. Emotion perception in emotionless face images suggests a norm-based representation. *Journal of Vision*, 9(1):1–11, 2009.

D. Neth and A. M. Martinez. A computational shape-based model of anger and sadness justifies a configural representation of faces. *Vision Research*, 50:1693–1711, 2010.

A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):107–119, 2000.

M. Prior. Is there an increase in the prevalence of autism spectrum disorders? *Journal of Paediatrics and Child Health*, 39:81–82, 2003.

G. Rhodes, S. Brennan, and S. Carey. Identification and ratings of caricatures: implications for mental representations of faces. *Cognitive Psychology*, 19:473–497, 1987.

S. Rivera and A. M. Martinez. Learning shape manifolds. *Pattern Recognition*, 45(4):1792–1801, 2012.

E. T. Rolls. A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion*, 4:161–190, 1990.

J. A. Russell. A circumplex model of affect. *J. Personality Social. Psych.*, 39:1161–1178, 1980.

J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110:145–172, 2003.

K.L. Schmidt and J.F. Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression. *Yearbook of Physical Anthropology*, 44:3–24, 2001.

L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Optical Soc. Am. A*, 4:519–524, 1987.

F.W. Smith and P.G. Schyns. Smile through your fear and sadness: Transmitting and identifying facial expression signals over a range of viewing distances. *Psychology Science*, 20(10):1202–1208, 2009.

J. Susskind, D. Lee, A. Cusi, R. Feinman, W. Grabski, and A.K. Anderson. Expressing fear enhances sensory acquisition. *Nature Neuroscience*, 11(7):843–850, 2008.

J.M. Susskind, G. Littlewort, M.S. Bartlett, and A.K. Anderson J. Movellanb. Human and computer recognition of facial expressions of emotion. *Neuropsychologia*, 45:152162, 2007.

F. Dela Torre and J. F. Cohn. Facial expression analysis. In Th. B. Moeslund, A. Hilton, V. Kruger, and L. Sigal, editors, *Guide to Visual Analysis of Humans: Looking at People*, pages 377–410. Springer, 2011.

M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3:71–86, 1991.

T. Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 43: 161–204, 1991.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, NY, 1998.

D. A. Wilbraham, J. C. Christensen, A. M. Martinez, and J. T. Todd. Can low level image differences account for the ability of human observers to discriminate facial identity? *Journal of Vision*, 8 (5):1–12, 2008.

R. B. Wilbur. Nonmanuals, semantic operators, domain marking, and the solution to two outstanding puzzles in asl. In *Nonmanuals in Sign Languages*. John Benjamins, 2011.

M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

A. W. Young, D. Hellawell, and D. C. Hay. Configurational information in face perception. *Perception*, 16(6):747–759, 1987.

L.A. Zebrowitz, M. Kikuchi, and J.M. Fellous. Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology*, 98(2): 175–189, 2010.