A Human-Markov Chain Monte Carlo Method For Investigating Facial Expression Categorization

Daniel McDuff (djmcduff@mit.edu)
MIT Media Laboratory
Cambridge, MA 02139 USA

Abstract

This paper demonstrates how a human-Markov Chain Monte Carlo (MCMC) method can be used to investigate models of facial expression categorization. Data were collected from four participants. At each step participants were asked to select a representation from a pair, that most resembled a particular emotional state; this was repeated iteratively. As such, they formed a component in the MCMC process. The representations were line drawn facial images with 10 nodes and four degrees of freedom. The judgements formed samples for a set of interleaved Markov Chains. These were mapped to a two-dimensional plane using Generalized Discriminant Analysis. We contrast the results of the MCMC task with those of a second discrimination task.

Estimates of the distributions along each of the four dimensions showed that for the outer eyebrow and lip corner variables one of the categories could be discriminated with confidence.

The average examples from both MCMC and discrimination tasks were both plausible. However, the MCMC method allowed for greater sampling from areas of high interest. Finally, we show that a naive Bayes classifier trained on the MCMC data can be used to successfully predict human classification in a discrimination task.

Keywords: MCMC; categorization; representations; facial expressions; emotion.

Introduction

The face provides an important channel for communicating affect. Much emotional information is encoded in people's facial expressions (Darwin, Ekman, & Prodger, 2002). However, affect label mapping from facial expressions is often difficult to define. In this paper we apply a Markov Chain Monte Carlo (MCMC) method (Neal, 1993) to investigate facial expression categorization. Using humans as components in a MCMC process we demonstrate how we can sample from cognitive representations of facial expressions.

MCMC is a sampling method that can be used to estimate probability density functions. A parameter space is searched via Markov Chains. The sampling procedure forms a chain that can be shown to tend to the correct distribution (Neal, 1993). In an environment where the distributions of interest are likely to occupy a small subspace only, MCMC can be an efficient sampling method.

Emotions are controversially defined. However, Ekman and Friesen's (Ekman & Friesen, 1978) set of six basic emotions are an accepted set of simple examples. These six are used as a starting point for our study: anger, disgust, fear, happiness, sadness and surprise.

This paper investigates how people map observed facial expressions to affect labels. Griesser et al. (Griesser, Cunningham, Wallraven, & Bulthoff, 2007) consider a psychophysical investigation of facial expressions. Scene parameters were systematically manipulated in order to investigate the importance of particular facial regions in expression recognition. Padgett (Padgett & Cottrell, 1997; Padgett, 1998) investigates representations of facial images for emotion classification. However, only 97 images are included in the data set. As a result there are a limited number of examples in a high dimensional space from which participants were forced choose one. Both these studies consider a pre-scripted set of stimuli and do not allow efficient exploration of each participant's psychological representations by allowing them to accept and reject samples based on how they fit with the category. Padgett represents human face judgements under multi-dimensional scaling (MDS). Such a method allows for a quantitative measure of similarity in the relationships between facial expressions.

This work considers human labels for expressions rather than the subjects state when displaying the emotion. It is important to consider that a persons evaluation of another affect given their facial expression may not be representative of their actual internal state.

Reasonable facial expressions for a particular emotion label are likely to occupy only a small subspace of the total space of possible expressions. This motivates the use of an MCMC method. MCMC allows regions within a facial action feature space to be populated with labels more efficiently that a discrimination task.

In particular, we investigate the significance of each feature dimension in the categories found. We estimate the density distributions for each category along each dimension. For a simple three category case considered, certain dimensions allow a particular category to be discriminated with confidence.

This is the first work I am aware of that models the relationship between emotional states and facial expressions drawn from continuous values within a multi-dimensional feature space. We allow the participants to navigate to an area of high association with the particular label and sample from this region more frequently (Neal, 1993). Representations are not limited by the number of examples in a data set but only by the ranges placed on the variables.

Related Work

Nosofsky's Generalized Context Model (GCM) of classification proposes that people represent categories by storing exemplars in memory (Nosofsky, 1986). The prototype theory assumes a category's mental representation is based on a prototypic exemplar (Dopkins & Gleason, 1997). In contrast, the exemplar theory assumes a set of exemplars are encoded

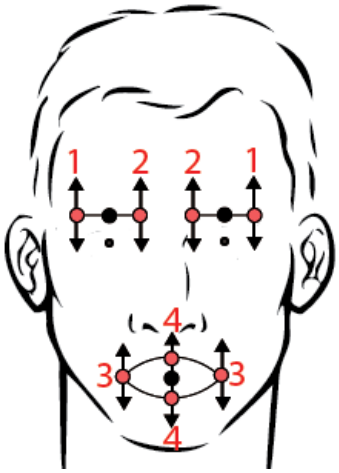

Figure 1: Face representation used in the tests. There are four degrees of freedom. 1. Position of outer eyebrows, 2. Position of inner eyebrows, 3. Position of lip corners and 4. Lip center separation. Center of the eyebrows was fixed (black node). Point about which lip center separation was measured was fixed (black node).

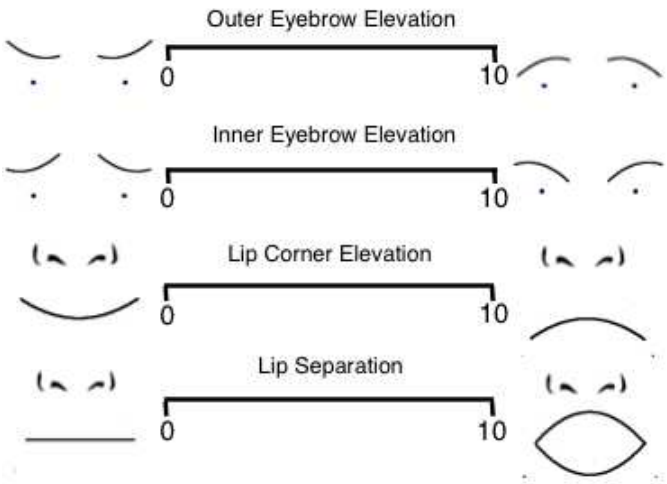

Figure 2: Continuous ranges of four free parameters on the face. Representations of the extreme cases are shown at either end of the scales.

in the category's mental representation (Nosofsky & Palmeri, 1997). A new entity is compared to the exemplars in order to establish whether it belongs to the category.

Sanbourn et al. (Sanborn & Griffiths, 2008; Sanborn, Griffiths, & Shiffrin, 2009) were the first to demonstrate the use of people as components in an MCMC algorithm, in order to explore psychological categories. A method was verified and used to demonstrate that human-MCMC can be used to estimate the structures of real-world animal shape categories.

Padgett (Padgett & Cottrell, 1997) considered representation of facial images for emotional classification. However this study is constrained by the fact that the facial image data set used was limited to a small number of images. The training data relied upon is limited in many cases as the images must be subject to agreement by expert labelers.

Methodology

This is the first investigation, to my knowledge, using cartoon representations of faces in order to investigate categorization of affect by facial expressions. As such it was necessary to begin with a facial representation having a small number of degrees of freedom. A cartoon representation was created with four degrees of freedom that allowed variation of eyebrows, lip corners and lip separation. These are demonstrated in Figure 1.

The limits placed on the displacement of each node are shown in Figure 2. The representation was symmetrical (eyebrows mirrored one another as did the left and right sides of the mouth). A restriction was applied in all tests that prevented the center of the eyebrows being the lowest point. This was the only restriction on the movement other than parameter range limits described. The degrees of freedom loosely correspond to the following action units which are identified in Ekman's (Ekman & Friesen, 1978) Facial Action-unit Coding System (FACS).

Outer Eyebrows - Outer Brow Raiser (AU2).
Inner Eyebrows - Inner Brow Raiser (AU1), Brow Lowerer (AU4).
Lip Corners - Lip Corner Puller (AU12), Lip Corner Depressor (AU15).
Lip Separation - Lips Part (AU25), Jaw Drop (AU26), Mouth Stretch (AU27).

In a set of initial tests two participants performed discrimination tasks with three facial representations. The first presented a mouth, nose and eyebrows where the nodes were joined by straight lines. The second added an outline of the face to the image. The third joined the nodes with smooth curves and also contained the outline of the face, as in Figure 1. The participants more consistently labeled the expressions given the third representation. As a result, this was used for the subsequent tests. This was a male face. Investigation into the effects of gender and ethnicity in this domain are not considered here.

All tests described in this paper were performed on a 15" MacBook Pro. Processing of the data and all GUI interfaces were created in MATLAB. None of the participants in the study were given rewards for completing the tasks. This study was approved by the Massachusetts Institute of Technology Committee On the Use of Humans as Experimental Subjects (COUHES).

Experiments

Three experiments were designed. The preliminary experiment was carried out to identify appropriate categories for the

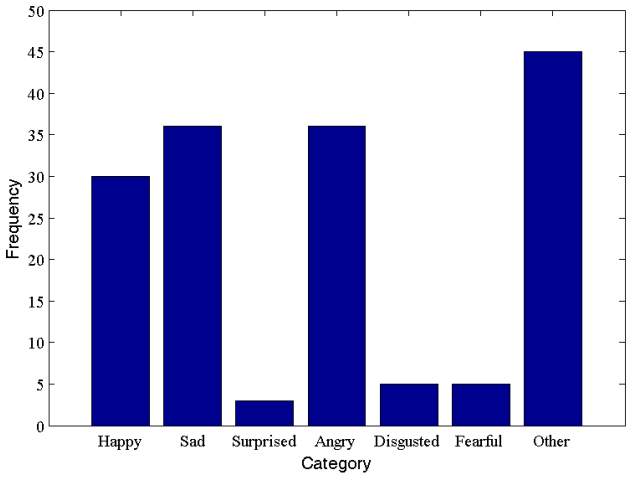

Figure 3: Histogram of results from the preliminary experiment, showing the frequency with which each category was chosen. Four participants labeled 40 different faces each.

human-MCMC tests. The human-MCMC experiment was then conducted to collect samples from these categories. The discrimination experiment was carried out to validate the distributions formed by the MCMC tests.

Preliminary Experiment

In a preliminary experiment four participants were separately shown a series of 40 cartoon faces and were asked to visually categorize them as angy, disgusted, fearful, happy, sad, surprised or other. The visual stimuli were generated from a uniform distribution over the parameter ranges shown in Figure 2. Representations outside these ranges were not considered as they were significantly different from natural movements, as judged by two participants in the initial tests.

Figure 3 shows a histogram of results from the preliminary discrimination experiment. Surprised, disgusted and fearful were each identified as the expression label in less than 5% of cases.

The results demonstrate that the four degree of freedom faces were not versatile enough to clearly represent all of the states. For instance the widening of the eyes that might be expected in a fearful expression was not represented.

There are likely to be many other indicators that influence our judgement of a person's affect that are not captured here. Ekman's facial action coding system (FACS) contains over 60 facial actions and movements many of which have been shown to discriminate between affective state (El Kaliouby & Robinson, 2005). These include skin texture changes, more subtle facial actions and movements. Examples are: nose wrinkles, head nods, shakes and tilts. Contextual information is also absent in our stimuli.

As a result, the affect categories were restricted to happy, sad and angry, which were the 3 most commonly identified categories in the preliminary experiment.

Human-Markov Chain Monte Carlo Experiment

Markov chain Monte Carlo (MCMC) is a sampling technique. At each step of the algorithm a proposed state is compared to the current state and one is rejected. The accepted state becomes the current state for the next step. The desired distribution is approximated using the Markov chain formed by the accepted samples. In this experiment, the MCMC analysis was performed by presenting two representations, one the current state in the chain and the other a proposed representation. The participants were asked: 'Which one is the more happy face?' for chain one, 'Which is the more sad face?' for chain two and 'Which is the more angry face?' for chain three. They selected the appropriate choice using a mouse click on a button below the appropriate picture.

Sanborn et al. identified in their human-MCMC analysis of animal representations that decision rule biases could form towards the current state or proposal (Sanborn et al., 2009). This led to unfavorable effects on the outcomes. In order to reduce the effect of such problems the MCMC chains for happy, sad and angry were interleaved. The decision to sample from a particular chain at any point was random and occurred with equal probability for all chains. As such, over many trials an approximately equal number of samples were taken from each category. The current and proposed states were displayed side by side on the screen during the tests.

Each of the MCMC chains was initialized by drawing a set of values from a uniform distribution over the lower 20% of the ranges in Figure 2. The proposed states were drawn from a multivariate Gaussian distribution with the current state as the mean and a diagonal covariance matrix. The standard deviation of the variables was set to 8% of their total range. In preliminary tests this was found to give a proposal acceptance rate from 30-50%. The ranges of the variables for the MCMC test are shown in Figure 2. If a proposal was outside the range then it was rejected and another set of samples taken.

Many studies fail to carefully consider the the impact of the experimental design on the data collected. To mitigate the effect of biases due to the participants not moving the cursor an unbiased coin flip was used to decide whether the current state would appear on the right or the left hand side of the screen. The select buttons were placed close together in order to minimize the effort required to change between the two.

Four participants performed the task. Participants 1, 2 and 3 evaluated 750 pairs over three chains and participant 4 evaluated 350 pairs over three chains, they all took between 30 and 60 minutes to complete the task. Table 1 shows the statistics from the MCMC experiment. The acceptance rate averaged over the whole participant pool was 36.5%.

In carrying out these tests we must be aware of assumptions made that may affect the results. Firstly, the MCMC method assumes that participants accept proposals by a rule that accepts less likely proposals with a certain probability. Secondly, the Markov assumption is that decisions are based on the current pair of stimuli. In such an experiment where the participants were each asked to evaluate a large number of

	No. of Samples			Acceptance %		
	Happy	Sad	Angry	Happy	Sad	Angry
P1	241	267	242	38	43	34
P2	231	271	248	53	41	37
P3	237	244	274	33	38	41
P4	113	114	123	30	20	30

Table 1: Participant's statistics. Number of samples per chain. Acceptance % per chain.

images they may make judgements based on previous images or may become bored with a particular image.

Discrimination Experiment

In this task the participants were presented with a single representation and asked to categorize it as happy, sad or angry. The representations were drawn from uniform distributions over the ranges shown in Figure 2. 750 different stimuli were categorized. The human-MCMC method allows sampling from the probability from the distribution in the the parameter space associated with each category. Thus even in the same context discrimination and MCMC would produce different information (Sanborn et al., 2009).

Results and Discussion

Human-MCMC is a sampling method. The data collected was in four dimensions (outer eyebrow, inner eyebrow, lip separation and lip corner dimensions). The samples obtained from the MCMC tests were mapped to a two dimensional plane that best discriminated between the expression distributions. This was carried out in order to create a visual structure of the expression categories (Olman & Kersten, 2004). The dimensionality reduction was performed using Generalized Discriminant Analysis (GDA) with a Gaussian kernel. GDA is a method of combining features so as to separate classes within the data. Figure 4 shows the resulting chains for all four participants. Using this visualization a judgement was made on how many samples should be rejected in order that the distributions were stationary. The number of samples burned (samples removed from the start of a chain) per chain was 40, leaving the average chain length 213 samples. The GDA was then performed on the samples in four dimensional space that remained after burn-in. Figure 5 shows the resulting samples for the four participants. The average faces for each participant and each category are shown in Figure 6. A mean face for each category, aggregated across the whole participant pool is shown in Figure 5. These faces appear to be reasonable examples of the three categories. This result in part supports the use of the MCMC method.

In these tasks, with only three categories in a limited dimensional space the categories can be separated effectively. However, if there were a great number of categories a Multi-Dimensional Scaling (MDS) representation could be created. We can calculate the similarity of categories by counting the confusions between pairs of stimuli (Rothkopf, 1957; Nosofsky, 1987). A potential downside of MDS is that it does not find an explicit mapping function from the parameter space. Sanborn et al. (Sanborn et al., 2009) use Dimensionality Reduction by Learning an Invariant Mapping (DrLIM) (Hadsell, Chopra, & LeCun, 2006) that does provide an explicit function. This was not tried here but would be worth considering in future work.

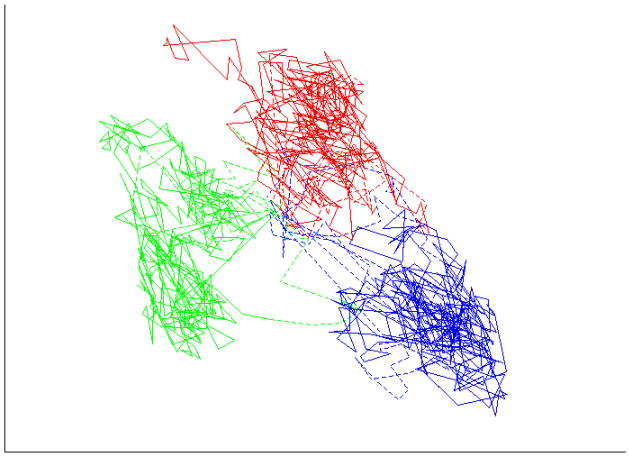

Figure 4: MCMC chains from all participants, before burn-in samples were removed, mapped to the plane that best discriminates between the categories. The dotted lines show the burn-in lengths chosen visually, the first 40 samples from each chain. Chain one - happy (green), chain two - sad (blue), chain three - angry (red).

Within a large parameter space the categories are likely to occupy small subspaces only. As a result a method such as MCMC that allows sampling from the whole parameter space but enables navigation to a particular region is useful compared to a discriminative test that samples from the space randomly.

However, in Figure 6 we compare the mean faces from the MCMC task and the discrimination task for one participant. In both cases the mean representations are reasonable examples. This suggests that the advantage of the MCMC method is not seen in this four dimensional space with the ranges described. As we increase the ranges and the number of degrees of freedom the space will increase greatly in size and it is likely that the benefit of the MCMC method will become apparent.

The discrimination experiment stimuli were categorized using the distributions found from the MCMC results. A naive Bayes classifier with Gaussian kernal was fitted to the four dimensional human-MCMC samples. Using this model the most likely label for each of the discrimination stimuli was chosen. These labels were then compared to the human responses.

The model matched the human identification of the stimuli in 70.1% of cases. This is much better than chance at 33%. The error is likely to be due to the fact that the discrimination

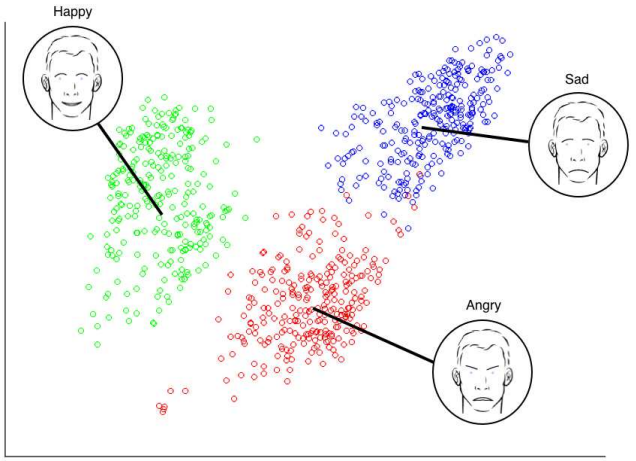

Figure 5: Scatter plot of samples from the four participants, after burn-in, mapped to the plane that best discriminates between the categories. The average face for each category is shown. Samples from: chain one - happy (green), chain two - sad (blue), chain three - angry (red).

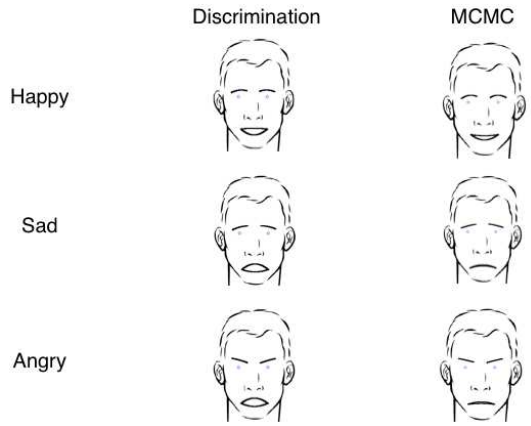

Figure 6: Comparison of mean faces for one participant in the discrimination task and MCMC task.

stimuli were generated from uniform distributions over the ranges. As such, many were far from the samples generated by the MCMC method. It is likely that many of the discrimination stimuli would not have been classified as any of the three categories if there had been other alternatives. Testing on results of a discrimination task with an 'other' option may produce even stronger performance.

For each of the dimensions the probability distributions for each category were estimated from the human-MCMC samples. The samples were separated into 25 equal size bins. Gaussian Process Regression (GPR)[1] was then used to approximate the distributions. A squared exponential (SE) covariance summed with an independent noise function was used. This does not make the assumption of an underlying structure but rather assumes the function is infinitely smooth. The characteristic noise scale and signal variance were set to one and the noise variance also to one. The hyper-parameters could be adjusted further. However, for a qualitative representation of the distributions given by the data these were reasonable choices.

Figure 7 shows the estimated density plots for each dimension after aggregating the data from all participants. It shows that in some dimensions (lip separation, inner eyebrow) none of the categories are significantly distinguished from the other two. However in the cases of the outer eyebrow and lip corner dimensions one of the categories was distinct. For the outer eyebrow dimension the distribution for anger is significantly different from the distributions for happy and sad. For the lip corner it is happy that is more distinguishable. The sad category distributions were not significantly different from both of the other two in any of the cases.

There are certain assumptions and limitations within the experiment that must be noted. As described above, when a proposal was outside the range set it was automatically rejected. In certain cases this rule was enforced and the distribution met one of the boundaries. This is not necessarily a negative point as the ranges restricted the participants to move within a space of reasonably natural expressions. We see from Figure 7 that for the inner eyebrows and lip corners the distributions did push up against the boundaries to a certain extent. This is something to consider in future work.

We should also note some general comments about aspects of the experimental set up. We must consider the impact of participants becoming bored during the experiment and selecting their response arbitrarily. Many samples were required in order to generate stationary distributions. Ways of minimizing the effects of boredom should be considered in future.

Conclusions

This paper demonstrates that human-MCMC methods can be used to gain insight into facial expression categorization using simple cartoon representations. We demonstrated that from 750 samples over three categories the method provides reasonable mean representations for each of the categories and reasonable distributions. By using GDA we were able to map the four dimensional points to a plane and after burn-in reveal three categories. The sad and angry chain samples were not separable in two dimensions. The happy chain samples were separable.

We also show estimates of the distributions for each of the categories along each of the four dimensions. This reveals that for the features tested the lip corner is the best discriminator for happy expressions and the outer eyebrow the strongest for angry expressions. The sad distributions were not distinguishable from both happy and angry distributions in any of the cases.

The mean faces generated by the human-MCMC and discrimination tasks were both reasonable and neither signifi-

[1] Rasmussen and William's GPML toolbox was used for this task.

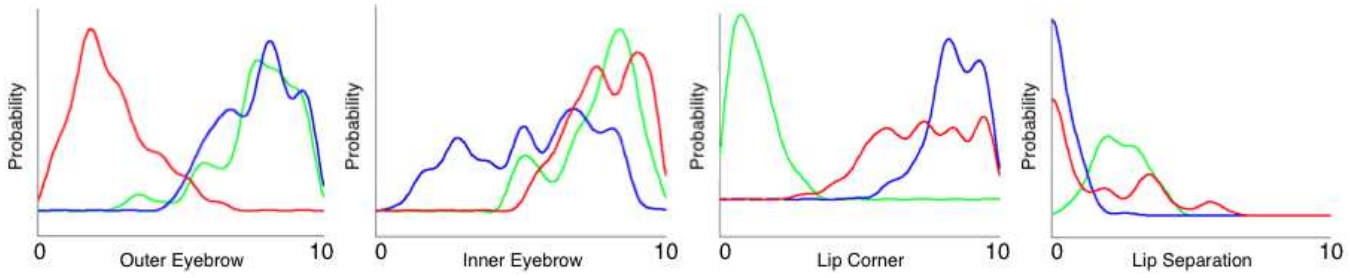

Figure 7: Density estimates for each of the four parameters aggregated over all the participants. The parameter dimensions correspond to the ranges shown in Figure 2. Chain one - happy (green), chain two - sad (blue), chain three - angry (red).

cantly more realistic than the other.

A naive Bayes classifier trained on the aggregated samples generated from the MCMC task performed strongly predicting over 70% of the human labels in the discrimination task correctly.

Further Work

This paper describes the first investigation evaluating human facial expression categorization using a human-MCMC method. It justifies a basis for applying a human-MCMC method for exploring people's representations of facial expressions. Griesser et al. (Griesser et al., 2007) demonstrate the use of detailed computer avatars that can realistically demonstrate skin texture changes as well as facial actions. This type of stimuli could be used in order to seriously investigate a wider range of categories. It would also allow more detailed investigation of the degree to which specific dimensions allow discrimination in terms of affect.

Sanborn et al. (Sanborn et al., 2009) suggest that the human-MCMC method may be used to test models of categorization. Prototype models produce unimodal distributions. Exemplar models are more flexible. As such it is difficult to establish whether a category distribution more closely resembles a prototype or exemplar model in many cases but rather we can test whether a distribution has properties that rule out a prototype model (Sanborn & Griffiths, 2008; Sanborn et al., 2009).

Acknowledgments

This work was funded by the MIT Media Lab Consortium.

References

Darwin, C., Ekman, P., & Prodger, P. (2002). *The expression of the emotions in man and animals*. Oxford University Press, USA.

Dopkins, S., & Gleason, T. (1997). Comparing exemplar and prototype models of categorization. *Canadian Journal of Experimental Psychology*, *51*(3), 212–230.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists.

El Kaliouby, R., & Robinson, P. (2005). Generalization of a vision-based computational model of mind-reading. *Proceedings of First International Conference on Affective Computing and Intelligent Interaction*, 582–589.

Griesser, R., Cunningham, D., Wallraven, C., & Bulthoff, H. (2007). Psychophysical investigation of facial expressions using computer animated faces. In *Proceedings of the 4th symposium on applied perception in graphics and visualization* (p. 18).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Proc. computer vision and pattern recognition conference (cvpr06)*.

Neal, R. (1993). *Probabilistic inference using Markov chain Monte Carlo methods*. Citeseer.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87–108.

Nosofsky, R., & Palmeri, T. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–299.

Olman, C., & Kersten, D. (2004). Classification objects, ideal observers & generative models. *Cognitive Science*, *28*(2), 227–239.

Padgett, C. (1998). *A neural network model for facial affect classification* (Tech. Rep.).

Padgett, C., & Cottrell, G. (1997). Representing face images for emotion classification. *Advances in neural information processing systems*, 894–900.

Rothkopf, E. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*(2), 94–101.

Sanborn, A., & Griffiths, T. (2008). Markov chain Monte Carlo with people. *Advances in neural information processing systems*, *20*.

Sanborn, A., Griffiths, T., & Shiffrin, R. (2009). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*.