Subjective Optimization

Chung-Cheng Chiu and Stacy Marsella

University of Southern California,
Institute for Creative Technologies,
12015 Waterfront Drive,
Playa Vista, CA 90094
{chiu,marsella}@ict.usc.edu

Abstract. An effective way to build a gesture generator is to apply machine learning algorithms to derive a model. In building such a gesture generator, a common approach involves collecting a set of human conversation data and training the model to fit the data. However, after training the gesture generator, what we are looking for is whether the generated gestures are natural instead of whether the generated gestures actually fit the training data. Thus, there is a gap between the training objective and the actual goal of the gesture generator. In this work we propose an approach that use human judgment of naturalness to optimize gesture generators. We take an important step towards our goal by performing a numerical experiment to assess the optimality of the proposed framework, and the experimental results show that the framework can effectively improve the generated gestures based on the simulated naturalness criterion.

1 Introduction

One of the main challenges in building a virtual human is to create its non-verbal behaviors such as gestures. Instead of spending time defining these animations manually, an alternative approach is to build a gesture generator to generate animations for dialogs automatically. One common design for building gesture generators is to apply machine learning algorithms to model the relationship between dialogs and gestures from human conversation data, and the derived model can then generate animations for similar dialogs automatically. The most common design for existing machine learning approaches trains models to fit the training data, but this conventional design is orthogonal to the actual goal of gesture generators, which is to produce natural motions that match well the corresponding dialogue. Although human conversation data collected for training are samples of natural gestures, unless the machine learning model can achieve 100% accuracy on fitting the data while also generalize to novel dialogue, there is no guarantee that the resulting model can generate natural motion. Conversely, since samples in the training data represent only a small portion of possible human gestures, there are many natural gestures that are quite different from

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 204–211, 2012.

the training samples that the gesture generator can learn to generate. Thus, there is a gap between the actual goal and existing approaches, such that the criterion of existing approaches makes the problem more challenging. One way to reduce this gap is to develop a naturalness criterion and use this criterion to optimize gesture generators. The naturalness criterion is a subjective criterion and a reasonable way to acquire this criterion is to collect feedback from humans.

The goal of this work is to propose an algorithm to refine gesture generators using the naturalness criterion. We adopt the idea of pairwise comparisons to help address the issue of the noisiness of absolute subjective judgments. To use these judgments to help learn a model of gesture generation, we frame the problem as a *dueling bandits problem* which optimizes the model from the results of the pairwise comparisons. At each step, the process generates a gradient vector to modify the parameters of the gesture generator and then generates gestures with the new parameters. If the generated animations are evaluated to be more natural than the original one, then the process generates a new set of parameters based on the gradient direction and use the parameters for the next optimization loop. The procedure performs several rounds to refine gesture generators and is applicable for gesture generators with numeric parameters. In this work we incorporate our framework into the gesture generator based on the Hierarchical Factored Conditional Restricted Boltzmann Machines (HFCRBMs) which generates motions based on given prosody.

Because HFCRBMs have very high dimensional parameters and optimizing the model with human subject evaluations is costly, in this paper we perform a numerical experiment to assess the optimality of the proposed framework. We define a metric to simulate human judgment and provide numerical evaluations for the optimization results. The numerical experiment shows that the algorithm can improve the gesture generator significantly.

2 Optimization Framework

In our optimization framework, the process first calculates a gradient vector for the parameters of the specified gesture generator and modifies the parameter with the gradient to get a new model. The framework then uses the two models to generate gesture animations for the same dialogue, pairs animations for the same dialogue from both models together into videos, and evaluates them in a pairwise comparison. The evaluated results are applied to generate the new gradient vector. The optimization process is shown in Figure 1.

One issue raised by the process is the evaluation of the naturalness criterion. Naturalness is a subjective opinion and it requires subject evaluations. Thus, the optimization process requires a mechanism to get evaluations for the quality of the generated gestures based on individual judgments. A conventional approach is to ask individuals to give absolute scores of the gestures (e.g. rate gestures from 1 to 10). However, individuals in general are poor at making absolute judgments as compared to discriminating judgements [6]. In fact, individuals make

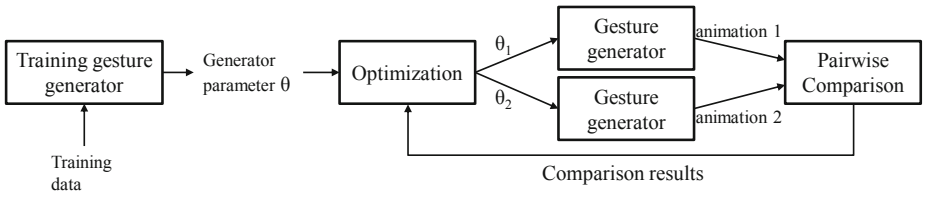

Fig. 1. The flow of the optimization framework

relative judgments for subjective evaluation and require reference points for making absolute judgments [7]. Since different individuals have different standards in mind, the evaluation results of absolute judgments can be inconsistent between individuals. Moreover, reference points of an individual change from one trial to the next [5] which suggests that the results of absolute judgments from the same individual may be inconsistent in itself. An alternative approach is to ask individuals to compare two animations and ask them to judge which one is better. This relative judgment approach not only resolves these biases but also makes the evaluation more reliable as individuals are more consistent in making qualitative judgments than estimating scores. Thus, our framework shows two animations to subjects and asks them to do pairwise comparisons to get naturalness evaluations.

Subject evaluations are very expensive, especially when each evaluation task takes several minutes. A common approach is to do crowdsourcing to collect a large number of evaluations with low cost [4], but this approach can still become expensive in our case when the model is high-dimensional as it may require hundreds of optimization iterations to get a reasonable improvement. We use HFCRBMs [1] as the model for gesture generators in the framework, and there are approximately one million parameters which requires a very large number of updates to optimize. Thus, in this work we define a metric as the naturalness criterion to bypass the expensive cost of human evaluation. We do not propose to replace human evaluations with the metric but rather use it here as an efficient approach to assess the framework. We currently use the metric in this work and will incorporate this with human evaluations as our next step. Thus, the framework is still designed for incorporating human evaluations.

The crucial step of the optimization process is to determine gradient vectors for model parameters. With the design of the pairwise comparisons, the step can naturally be formulated as a *dueling bandits problem* [12]. The dueling bandits problem refers to finding an optimal decision that minimizes regret based on pairwise comparisons. Its discrete version is the K-armed dueling bandits problem [11] in which there are a collection of K bandits and the process needs to determine which bandit leads to the best result based on pairwise comparisons. Dueling bandits problems, in contrast, do not feature K distinct choices of bandits but optimize over an entire finite continuous space of them. Each iteration comprises a comparison between two selected bandits A and A', and the optimization process uses the comparison result to determine the two bandits for

Algorithm 1. Optimization Algorithm

Input: $\gamma, \delta, \theta, X$
  $G \leftarrow generate(\theta, X)$
  for $t = 1 \rightarrow T$ **do**
    $u \leftarrow$ Sample unit vector uniformly with the same dimension as θ
    $\theta' \leftarrow \theta + \delta u$
    $G' \leftarrow generate(\theta', X)$
    if Comparisons show G' is better than G **then**
      $\theta \leftarrow \theta + \gamma u$
      $G \leftarrow generate(\theta, X)$
    end if
  end for

the next comparison. The evaluation results derived from pairwise comparisons are assumed to be noisy where there is a chance that A is evaluated to be better than A' despite A' actually being better than A. The optimality of the selected bandits is quantified as a regret, defined as:

$$R_T = \sum_{t=1}^{T} \epsilon(A_t^*, A_t) + \epsilon(A_t^*, A_t'),$$

where A_t and A_t' are two bandits selected at time t, and A_t^* is the best bandit at time t.

Dueling bandits problems can naturally be related to optimization problems where bandits correspond to gradient vectors for model parameters, and the decision process is equiavlent to choosing an effective gradient for optimizing the model. An algorithm called *Dueling Bandit Gradient Descent* (DBGD) has been proposed to perform online optimization with pairwise comparisons, and its regret bound has been shown in previous work [12].

We apply DBGD to perform online optimization for gesture generators. The algorithm is described in Algorithm 1. In Algorithm 1 δ is the step size for exploring effective gradient direction, and γ is the step size for exploiting this gradient direction in the current model, θ denotes parameters of the gesture generator, X denotes required data for generating gesture animation, and function *generate* represents the specified gesture generator which generates animations Gs based on θ and X. Previous work [12] has shown that when G' is evaluated as better than G the corresponding u indicates a gradient direction that can improve the model.

The algorithm takes θ as input and uses it as an initial point of the optimization process. The optimization process depends on the naturalness evaluation, and without initializing the parameter with the original training algorithm, the generated animations will in general be too poor to make such a comparison. Thus, our framework trains gesture generators with their original training algorithm before starting the optimization process. This algorithm is applicable for gesture generators with numeric parameters.

2.1 Gesture Generator

We follow the same design of the previous work [1] for our HFCRBM-based gesture generator. The HFCRBM-based gesture generator generates gesture motion based on prosody information and past motion. It learns the model from motion capture data of human conversations. The HFCRBM decomposes the gesture learning problem into two parts: it first learns the hidden factors of generating human motion and then learns the correlation between prosody and these hidden factors. This design is based on the idea that human motion is driven by a set of motor signals that in combination produce a sequence of movements. In contrast, the motion capture data is represented as real valued vectors that in essence obscure the factors and constraints that were involved in generating the data. Thus, the model first infers the underlying causes of motion and then learns the relation between speech and the hidden factors. The HFCRBM is comprised of two components, a reduced conditional restricted Boltzmann machine (RCRBM) [2] for inferring hidden factors of motion and a factored conditional restricted Boltzmann machine (FCRBM) [8] for modeling the relationship between prosody and hidden factors. To learn gesture generators from human conversation data, the model first trains RCRBMs with motion capture data. After this training step, the HFCRBM maps motion data onto hidden factors with RCRBMs and trains the top-layer FCRBMs conditioned on audio features. For gesture generation with HFCRBMs, the generation process works by taking previous motion frames and audio features as input to generate the next motion frame. The generation of a sequence of motion is done in a recurrent way in that the generated motion frame becomes part of the input of the next generation step. The process can generate a motion sequence with the same length as the audio.

The HFCRBM for the gesture generator has many parameters, and performing our optimization algorithm with all the parameters is inefficient. It is more reasonable to pick a subset of parameters for the optimization process. We can use the fact that the HFCRBM comprises two separate modules and focus on optimizing only one of them. The reason for proposing this optimization framework is to improve the generalization of gesture generators for novel prosody, and therefore it is more reasonable to focus on improving the modeling among prosody and hidden factors with the naturalness criterion. Thus, we only update the parameters of FCRBMs. Because we only optimize the top model, in order to give it better control we choose to use RCRBMs as the bottom model instead of CRBMs [9], which results in a HFCRBM with the same architecture as [2]. In our HFCRBM model, the hidden layers of the RCRBMs and FCRBMs each have 300 nodes. The prosodic features for gesture generators at each time frame have a window of $\pm 1/3$ seconds.

2.2 Simulating Pairwise Comparisons

The crucial step of the optimization algorithm is the pairwise comparison. We define a metric to simulate human evaluation and simulate the pairwise comparison

by comparing the metric values of two animations. The metric function for an animation set G is defined as:

$$metric(G) = \sum_{i=1}^{n} ||cor(g_i^*) - cor(g_i)||_2^2 / 2n.$$

In this function $G^* = g_1^*, ..., g_n^*$ represents a set of human motion for the corresponding prosody, and $G = g_1, ..., g_n$. The function $cor(g)$ is defined as:

Input: g
$p \leftarrow pitch(g)$
$i \leftarrow intensity(g)$
$v \leftarrow velocity(g)$
$a \leftarrow acceleration(g)$
return $correlation(p, a)$,$correlation(i, v)$

where $correlation(x, y)$ calculates the linear correlation of the two input sequences x and y. We proposed this metric under an assumption that the pitch is more likely to be correlated with the movement acceleration and the intensity is more likely to be correlated with the movement velocity. One possible alternative is to directly compare the generated motion and the motion capture data, but since the motion capture data only shows one instance of possible gestures this explicit comparison can limit the variety of generated gestures. Thus, we choose a more implicit comparison metric to increase the flexibility for the gesture generation.

3 Experiments

We assessed our framework by conducting an experiment to analyze its performance empirically. To train the gesture generator, we used a dataset originally created for a different study that examined how audio and body motion affected the perception of virtual conversations [3]. We followed the approach suggested in [1] to extract the data. Unlike their configuration, we only use prosody features as contextual information and exclude correlation parameters, as they did not seem to improve the quality of generated gestures. There are a total of 1140 frames (38 seconds) of training data.

We trained the gesture generator with the HFCRBM training process, and then used our optimization framework with the simulated pairwise comparison described in subsection 2.2. We used the training data applied in the HFCRBM training process for the simulated comparison process, in which the gesture generator is requested to generate gestures with the prosody in the training data and the generated gestures are compared with the corresponding motion capture data based on our defined metric. We ran the optimization loop for 1000 iterations, and the output of the simulation metric (can be understood as error values) of each successive training iteration are shown in Figure. 2. After 1000 iterations, the value drops to less than half of the original value.

The experiment shows that the algorithm can effectively improve the generation result of the gesture generator. One limitation of this framework for gesture

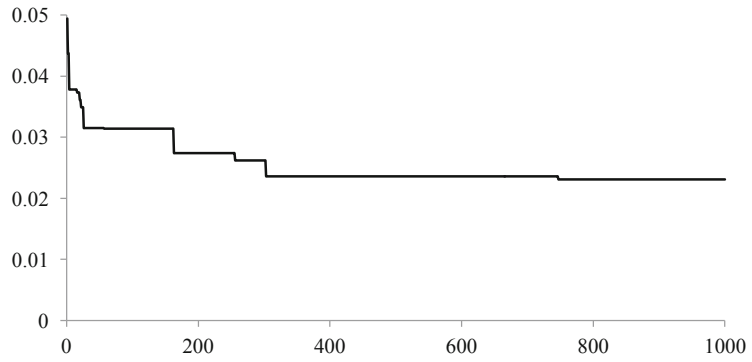

Fig. 2. Error values of the optimization process

generator is that when the model has many parameters, it requires a lot of iterations for the optimization process to improve the results. For example, [10] uses 10^5 iterations to train their model for search ranking. Also, there is no guarantee that the function we are optimizing is convex, which makes it necessary to try several initializations to avoid local minimum. These two properties suggest that we need a large number of pairwise comparisons to get promising results which leads to expensive demands on human effort. A possible solution is to improve the HFCRBM so that it can maintain similar quality in gesture generation by using far fewer parameters. Another method is to introduce some heuristic function based on domain knowledge to narrow the search space. Further work needs to be done before we can investigate the incorporation of real human judgment into this framework.

4 Conclusions

This work seeks to address the common problem of existing machine learning based gesture generators in which the training objective does not match the naturalness criterion people expect for gesture generators. We have proposed a framework to improve gesture generators with a naturalness criterion. The framework lays a foundation for training gesture generators using a naturalness criterion. Specifically, we applied our framework to improve a HFCRBM-based gesture generator. The framework identifies a gradient that can improve the model and updates the parameters iteratively. The optimization algorithm uses the information from comparing two generated gesture animations, and the pairwise comparisons are simulated with a metric based on prosody and motion. The efficacy of the framework is demonstrated in experiments that show significant improvement of the HFCRBM-based gesture generator. The major limitation of the framework is that the cost of the optimization process can become impractical when applied to gesture generators with too many parameters. Future work needs to address this parameter problem before we can proceed with moving from a simulated human judgments to actual human judgments.

Acknowledgements. This research was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) Simulation Training and Technology Center (STTC). The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Chiu, C.-C., Marsella, S.: How to Train Your Avatar: A Data Driven Approach to Gesture Generation. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 127–140. Springer, Heidelberg (2011)
2. Chiu, C.-C., Marsella, S.: A style controller for generating virtual human behaviors. In: Proceedings of the 10th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2011, vol. 1 (2011)
3. Ennis, C., McDonnell, R., O'Sullivan, C.: Seeing is believing: body motion dominates in multisensory conversations. In: ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010, pp. 91:1–91:9. ACM, New York (2010)
4. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 453–456 (2008)
5. Mozer, M., Pashler, H., Wilder, M., Lindsey, R., Jones, M., Jones, M.: Improving human judgments by decontaminating sequential dependencies. In: Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 23, pp. 1705–1713 (2010)
6. Stewart, N., Brown, G.D.A., Chater, N.: Absolute identification by relative judgment. Psychological Review 112(4), 881–911 (2005)
7. Stewart, N., Chater, N., Brown, G.D.: Decision by sampling. Cognitive Psychology 53(1), 1–26 (2006)
8. Taylor, G., Hinton, G.: Factored conditional restricted Boltzmann machines for modeling motion style. In: Bottou, L., Littman, M. (eds.) Proceedings of the 26th International Conference on Machine Learning, pp. 1025–1032. Omnipress, Montreal (2009)
9. Taylor, G.W., Hinton, G.E., Roweis, S.T.: Modeling human motion using binary latent variables. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, vol. 19, pp. 1345–1352. MIT Press, Cambridge (2007)
10. Yue, Y.: New learning frameworks for information retrieval. Ph.D. thesis, Cornell University (2011)
11. Yue, Y., Broder, J., Kleinberg, R., Joachims, T.: The k-armed dueling bandits problem. In: Conference on Learning Theory (2009)
12. Yue, Y., Joachims, T.: Interactively optimizing information retrieval systems as a dueling bandits problem. In: ICML (2009)