A MARKOV CLUSTERING METHOD FOR ANALYZING MOVEMENT TRAJECTORIES

Jacob Goldberger[1], *Keren Erez*[2], *Moshe Abeles*[2]

[1]Bar-Ilan University, School of Engineering, Ramat-Gan 52900 Israel
[2]Bar-Ilan University, Brain Research Center, Ramat-Gan 52900, Israel

ABSTRACT

In this study we analyze monkeys' hand movement; our strategy is compositionality, division of complex movement into basic simple components-primitives. Representing each trajectory segment as vectors of directions, we model the movement trajectory as a large Markov process where each state is related with an average trajectory pattern. In the next step, in order to find the movements primitives, we cluster the Markov states according to their probabilistic similarity. We present an information theoretic co-clustering algorithm which can be interpreted as a block-matrix approximation of the Markov transition matrix. The performance of the suggested approach is demonstrated on real recorded data.

1. INTRODUCTION

The idea of complex constructs that can be divided into small units, and then subdivided into subunits (and so forth) can be referred to as a compositionality. Compositionality, as a way of thinking, may give us tools to better understand, and maybe predict these constructs' future form. Our research deals with movement trajectories and the ways in which they can be viewed as a compositional constructs. Stated differently, we aim to find the basic rules that shape the movement trajectories.

Past research observed subjects moving their hand between pairs of targets and found that they tend to generate straight hand paths with single-peaked, bell-shaped velocity profiles. These, so called stereotyped movements, are invariant even after changing the size, rotation, translation and temporal scaling [7]. Present studies regard these basic movements as "primitives". Many studies claim that complex movements are composed of basic building blocks thus adding more support to the notion of primitives. These building blocks can be executed sequentially or concurrently (through parameterized superposition), to create a large movement repertoire [12] . The debate over the existence of primitives was quite fertile over the years. One of the first findings was that it only takes a dozen primitives to encode a frog's entire motor repertoire [2]. It was found that if the movement target was shifted during movement, the arm trajectory would become curved in such a way that involved a vectorial summation of two basic movement elements. The summed trajectory was then smoothness as can be articulated by a cost function [5]. Studies have shown that the frog and the rat's complex limb movements may be generated by a vectorial summation of modular force fields in the spinal cord [3, 14]. These findings may lead to the notion that modular primitives are part of both the planning and the execution of multi-joint limb movements. Another study showed that one year old children's movement can be decomposed into a sequence of stereotyped movements each resembling simple basic movements of adults [9, 10, 1]. It was also found that stroke patients' initial movements could be easily decomposed and showed invariant velocity profiles [11]. Recently, Flash and Hochner [6] have published a review, describing motor primitives and compositionality. They defined motor and movement primitives, discussed the nature of these primitives, their internal representation, and the rules governing the generation of a large repertoire of movements from a limited set of elements.

Here we analyze monkeys' hand movement; our strategy is compositionality, division of complex movement into basic simple components - primitives. We search for primitives by the following strategy; first, we divide the drawing movements to many small strokes. We then group these strokes according to the set of directions they take. Then we compute the transition probability between groups - thus treating the drawing as a Markov process and the group as states of this process. These states may be regarded as "phones" in speech; we then try to combine several phones into equivalent clusters of states (the clusters of states may be regarded as "phonemes" of speech). These clusters of states (phonemes) may be treated as the primitives from which the more complex drawing is composed. Combining states into clusters is based on the idea that such combination is useful as long as there is not much loss of information in the Markovian transition matrix.

The paper proceeds as follows. In Sections 2 and 3 we formally present the Markov clustering problem and propose an efficiently computed algorithm. Related work is discussed in Section 4. Section 5 describes the pre-processing that was applied on the hand-movement data following by experimental results of the clustering algorithm.

2. NOTATION AND MODEL

The process of forming primitives of hand movements, via combining groups of similar-strokes into clusters according to their behavior along the time axis, is translated in our approach into the mathematical problem of clustering the states of a Markov process. In this study we propose an efficient information-theoretic clustering algorithm for this clustering task.

Let $X = \{x_t\}$ be an n-valued stationary first order ergodic Markov process defined by the $n \times n$ transition matrix A where $A_{ij} = p(x_1 = j|x_0 = i)$. A function $\pi : \{1, ..., n\} \to \{1, ..., m\}$ defines a partition of the state-space into m subsets $w = \{w_1, ..., w_m\}$ such that $w_k = \pi^{-1}(k), \quad k = 1, ..., m$. Utilizing π we can define a new m-state lumped Markov process $Y = \{y_t\}$ in the following way. Define $(y_0, y_1) = (\pi(x_0), \pi(x_1))$, i.e. the joint distribution of y_0 and y_1 is defined via the following Markov chain relation:

$$y_0 \xleftarrow{\pi} x_0 \xrightarrow{A} x_1 \xrightarrow{\pi} y_1$$

More explicitly, the joint distribution of y_0 and y_1 is:

$$p(y_0 = k, y_1 = l) = \sum_{i \in w_k} \sum_{j \in w_l} p(x_0 = i, x_1 = j)$$

Since x is a stationary process, it can be easily verified that the marginal distributions of y_0 and y_1 are the same. Let Y be the stationary m-state Markov process defined by the stochastic matrix $p(y_1|y_0)$.

We want to cluster the states of the process X to form a new reduced Markov process that best preserves the structure of the original process X. Intuitively, states i and j are viewed as similar if both the future conditional distributions $p(x_1|x_0 = i)$ and $p(x_1|x_0 = j)$ and the past conditional distributions $p(x_0|x_1 = i)$ and $p(x_0|x_1 = j)$ are similar. The Information-Bottleneck (IB) principle [17] can be used to formalize this intuition. The IB principle for this case states that the best clustering function π of the n states into m clusters is the one that maximizes the mutual information $I(\pi(x_0); \pi(x_1)) = I(y_0; y_1)$ over all the partitions of the state-space into m subsets. Utilizing standard information-theory manipulation we can derive several equivalent forms for the cost function we want to minimize.

$$\begin{aligned} C(\pi) &= I(x_0; x_1) - I(y_0; y_1) \qquad (1) \\ &= D(p(x_1|x_0)\|p(x_1|y_0)) + D(p(y_0|x_1)\|p(y_0|y_1)) \\ &= D(p(x_0, x_1)\|p(y_0, y_1)p(x_0|y_0)p(x_1|y_1)) \\ &= H(y_0, y_1) + H(x_0|y_0) + H(x_1|y_1) - H(x_0, x_1) \end{aligned}$$

where $y_0 = \pi(x_0)$, $y_1 = \pi(x_1)$, D is the Kullback-Leibler divergence and H is the entropy function. The optimal state-clustering is the one that minimizes the information-loss function $C(\pi)$.

Before proceeding to the clustering algorithm we want to clarify a theoretical point. Note that given π there is yet another stochastic process that can be constructed. We can utilize π to define a Hidden-Markov Model (HMM) $Z = \pi(X) = \{\pi(x_0), \pi(x_1), \pi(x_2), ...\}$. Although the joint distributions of (y_0, y_1) and (z_0, z_1) are the same, generally the distributions of (y_0, y_1, y_2) and (z_0, z_1, z_2) are different and Z is even not necessarily Markovian. To clarify the relations between the processes Y and Z we need the following definition.

Definition: A Markov process X is weakly-lumped with respect to a partition w if $I(\pi(x_1); x_0|\pi(x_0)) = 0$, i.e. for each two subsets $w_k, w_l \in w$ the probability $p(x_1 \in w_l|x_0 = i)$ is constant over all $i \in w_k$.

A classical theorem states that X is weakly-lumped with respect to the partition w if-and-only-if $Z = \pi(X)$ is a Markovian process. In this case Z coincides with the process Y defined in the previous paragraph. The transition matrix of a weakly-lumped Markov process has a natural block-matrix structure that is imposed by the partition w. In our setup we are concentrating in Y which is always Markovian by definition.

3. THE CLUSTERING ALGORITHM

There is no closed-form solution for the optimization problem posed in Section 2. Several standard optimization algorithms can be utilized to find the best clustering. We can use agglomerative algorithm which is based on bottom-up merging procedure. We can use a K-means clustering algorithms where the Bregman distance is the Kullback-Leibler divergence. Alternatively we can apply a greedy sequential algorithm that can be viewed as a sequential version of the K-means algorithm. In this study we apply the sequential greedy algorithm which was found to perform well in terms of both clustering quality and computational complexity. The sequential clustering algorithm starts with a random clustering of the states into m clusters. We go over the n original states in a circular manner and for each state we check if moving it from one cluster to another can increase the cost function $I(y_0; y_1)$.

The basic step in each of the mentioned algorithms is composed of computing the distance function between two clusters. It can be verified that in our case the information-bottleneck principle implies that this distance is the information-loss caused by merging the two clusters into a single one, i.e. the difference between mutual-information of the reduced markov processes before and after the two clusters are merged. We derive next an explicit and efficiently computed expression for this distance between clusters.

Assume we are given a partition of the Markov states $w = \{w_1, ..., w_m\}$ and we want to compute the information loss caused by merging the clusters w_1 and w_2 to obtain

a new reduced partition $w' = \{w_1 \cup w_2, w_3, ..., w_m\}$ into $m-1$ groups. Let y_0 and y_1 be the Markov chain variables defined by w and let y_0' and y_1' be the Markov chain variables defined by w'. The information-loss can be efficiently computed in the following way.

$$d(w_1, w_2) = I(y_0; y_1) - I(y_0'; y_1') = \quad (2)$$
$$I(y_0; y_1) - I(y_0; y_1') + I(y_0; y_1') - I(y_0'; y_1') =$$
$$\sum_{i=1,2} p(w_i) D(p(y_0|y_1 = w_i) || p(y_0|y_1 \in w_{12})) +$$
$$\sum_{i=1,2} p(w_i) D(p(y_1'|y_0 = w_i) || p(y_1'|y_0 \in w_{12})) =$$
$$p(w_{12})(JS(p_{1|0}(\cdot|w_1), p_{1|0}(\cdot|w_2)) + \quad (3)$$
$$JS(p_{0|1}(\cdot|w_1), p_{0|1}(\cdot|w_2))) - p_{01}(w_{12}, w_{12}) I_{12}$$

where $w_{12} = \{w_1, w_2\}$ and $p_{01}, p_{1|0}, p_{1|0}$ are the joint and conditional distributions of the reduced-Markov random-variables y_0 and y_1. JS is the Jensen-Shannon divergence. $I_{12} = I(y_0; y_1|y_0, y_1 \in w_{12})$ is the mutual-information of the following joint-distribution matrix:

$$\frac{1}{p_{01}(w_{12}, w_{12})} \begin{pmatrix} p_{01}(w_1, w_1) & p_{01}(w_1, w_2) \\ p_{01}(w_2, w_1) & p_{01}(w_2, w_2) \end{pmatrix} \quad (4)$$

Hence, the distance measure $d(w_1, w_2)$, takes into account both the future and past conditional distributions. The possible overlap between these two distance components is subtracted from the sum. The sequential clustering algorithm requires the computation of the change in the cost function when moving a state from one cluster to another which can be efficiently done using expression (3).

One drawback of the sequential algorithm (in contrast to agglomerative approaches) is that the number of clusters should be given as an input to the algorithm. We can slightly modify the algorithm is such a way that we should just provide a rough estimation (upper bound) on the number of desired clusters. Consider the case of a cluster that contains a single object s. The iterative-sequential algorithm will not merge s into any other cluster because obviously this can not increase the cost function $I(y_0; y_1)$. The algorithm will always prefer to leave s as a single member of cluster. In the modified version we enforce a singleton cluster to be merged into another cluster. This step reduces the number of clusters by one. Utilizing this scheme, the number of output clusters can be adapted to the data. Note that in this method each of the output clusters should contain at least two members. The algorithm is summarized in Table 1. Since there is no guarantee that the algorithm finds the global optimum, we can apply the algorithm on several random partitions and choose the best local optimum.

Input: A Markov transition matrix $n \times n$.
Output: A partition of the Markov states into (at most) m clusters.

Algorithm:

1. Choose a random partition of the Markov state into m clusters.
2. Loop until there is no change
   - for $s = 1, ..., n$
     - Remove state s from its current cluster.
     - If s is the only member of its cluster, delete the cluster.
     - Merge s into the cluster w_k that minimizes the distance $d(\{s\}, w_k)$.

Table 1. The Markov-states clustering algorithm

4. RELATED WORK

Information-theoretic approaches have been intensively used for clustering and co-clustering methods [17, 4]. Unlike previous works, in our setup the same clustering function π is simultaneously applied to the two random variables x_0 and x_1. Ge et al. [8] considered the Markov-state clustering problem as a parameter estimation problem of a HMM. They viewed the reduced-state model as a constraint HMM. The original Markov-process is viewed as the observed part of the HMM and the constraint is that each observed symbol can appear only in one hidden-state. The relation between this approach and ours is related to the analogy between EM and IB described in [15]. Note that in our derivation (unlike similar approaches [8]) there is no need to recompute the entire score (whose computational complexity is $O(n^2)$) for each sequential update. The distance measure $d(w_1, w_2)$ is only based on a small part of the transition matrix related to the states in $w_1 \cup w_2$ and it can be computed in $O(n)$ operations. Meila and Shi [13] showed the connection between spectral clustering and clustering the states of a Markov transition matrix of a random-walk process defined by the pairwise-distance matrix. In their approach a cluster of states in characterized as follows. Once the process is in one of the members of the cluster in tends to remain in the cluster.

5. EXPERIMENTS

5.1. Data Pre-Processing

Hand movement analysis is our research main goal; our tools provide us ways to examine monkeys' hand move-

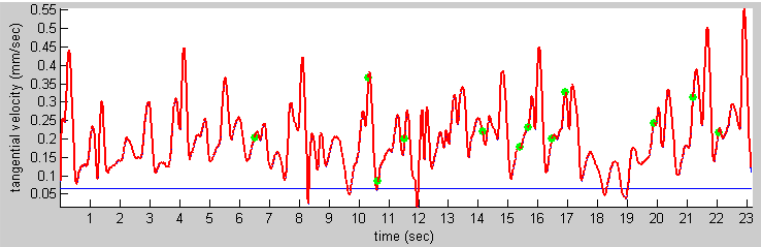


Fig. 2. Hand movement's tangential velocity as a function of time. The green dots represent the time at which reward was given. This graph depicts 23 seconds taken from a much longer session.

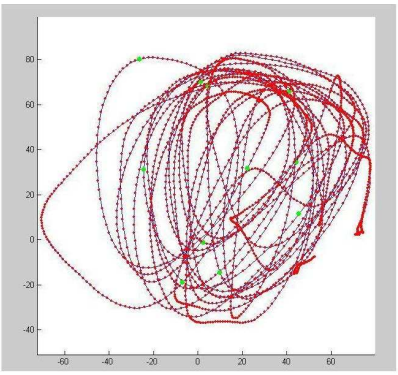


Fig. 1. A print of the trajectory of hand-location as a function of time.

ment. A monkey was trained to move his hand freely in a continuous two dimensional motion. Throughout the entire session the monkey held a manipulandrum, which monitored the hand exact location at all times on the workspace. The workspace was divided into a grid of 19 invisible hexagons. One of the hexagons was randomly selected to be the rewarded hexagon so that once the monkey moved the manipulandrum's cursor into that hexagon, it received a juice reward. After receiving the reward, the rewarded hexagon was replaced by another hexagon until the next reward. This process repeated itself several times. During the experiment, the manipulandrum angular position was transformed into X-Y coordinate and was recorded at a rate of 100Hz. We defined the movement period as a period of continuous motion with a speed exceeding a threshold of 2 cm/sec. The cursor's position was monitored and traced so that the outcome looked like scribbled circles drawn one on top of the other. Figure 1 shows a typical example of a print of the trajectory of hand-location as a function of time.

Using the speed profile, we segmented the trajectory so that each segment's initial velocity would correspond to a local minimum or maximum velocity along the entire trajectory. Naturally, every segment's final velocity corresponds to the next local maximum or minimum velocity along the entire trajectory. Note that the beginning and the ending points of the movement along the velocity profile, were regarded as local minimum and local maximum even though mathematically it has not been quite so. We used the local extrema in the tangential velocity as it was found that these points may be the points at which the movement speed parameters may abruptly change [18]. The hand movement's tangential velocity as a function of time is demonstrated in Figure 2.

Next we extracted features from each segment. Motivated by methods that were successfully applied to on-line hand-writing recognition systems (see e.g. [16]), we modeled the segments using angles of equidistant (in their time duration or length) sub-vectors. Note that angle parameter is invariant to both size and translation of the input segment. We divided each segment into ten equal fractions. As a result we obtained ten tangents, and more importantly, ten angles (made by the tangents and the horizon) as a reproducible description of the segment. Thus, each segment was represented as a single point in a 10-dimensional feature space.

We utilized the Mixture of Gaussians model (MoG) to cluster segments into groups based on the angle-vectors representation. For simplicity (and to avoid over fitting), we assumed that (within a group) every one of the 10 angles in each segment was sampled independently from the rest, even though it is clear that smooth hand movement does rely on some dependency among adjacent angles. This assumption is translated into a diagonal structure that is imposed on the covariance matrix of each Gaussian component. The EM algorithm was utilized to find the maximum-likelihood model parameters. While clustering the feature vectors into clusters we had to consider the cyclic nature of angles. Hence, angles such as 1 degree and 359 degrees should be considered as close to one another. While computing the Gaussian density of a segment we have always computed the distance between the mean angle and the ob-

served angle along the direction that yields a smaller distance. Since all the variances were found to be much smaller than 360, this approximation makes sense.

As a result of the EM learning step, we obtained a MoG model composed of a large number of components. Given the MoG model, we labeled each of the segments to one of the Gaussian component based on maximum-posterior probability. The labels were used to estimate a Markov transition matrix. Note that alternatively we can utilize a Hidden Markov Model (HMM) to estimate simultaneously both the Gaussian emission distributions of the segments and the hidden Markovian structure. However, since the multi-Gaussian emission distribution is much more dominant than the Markovian transition probabilities, using HMM was resulted with very similar parameter set. The pre-processing step, therefore, transforms the hand-movement segments into a large Markov transition matrix.

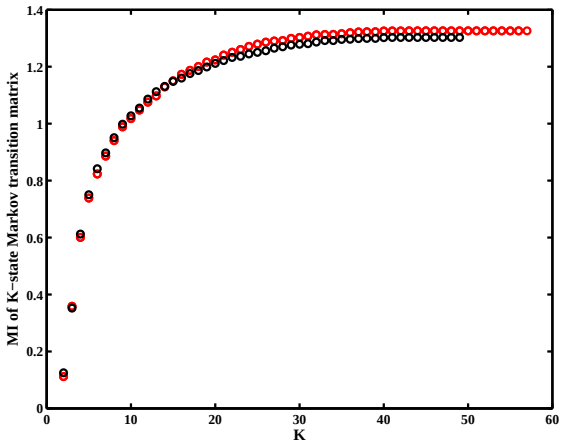


Fig. 3. The mutual-information (MI) score of the clustered Markov process as a function of the number of clusters. The black circles indicate MI values as a based on equal-time division features. The red circles indicate MI values as a function of equal-distance division features.

5.2. Clustering Results

Our next step is grouping the states of the Markovian process into clusters. The clustering algorithm, described in Section 3 was utilized for that task. Figure 3 presents the mutual-information (MI) clustering-quality score, $I(y_0; y_1)$, as a function of the number of clusters. As can be seen, we obtained a "knee" typed curves that is, range of values at which a significant decrease in the MI took place. We implemented the clustering algorithm upon those segments that were assigned to specific state based on equal time division or based on equal length duration. As can be seen in

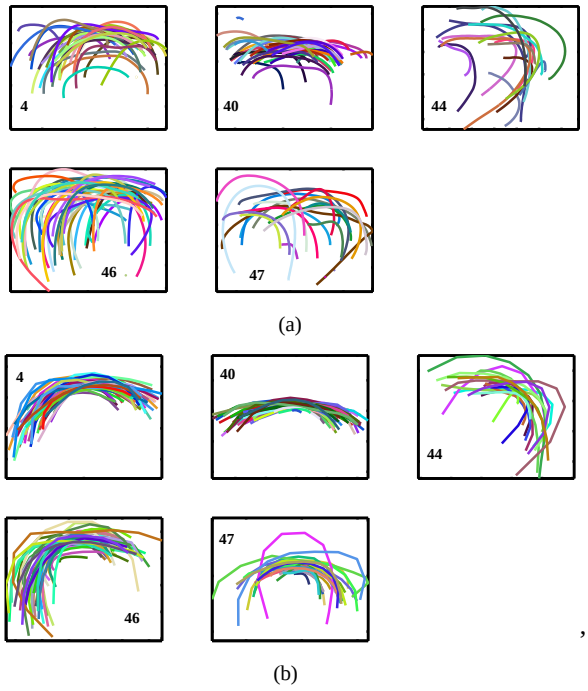


Fig. 4. One of the clusters of markov-state, based on equal-time division. The hand-movements segments in each state (plotted in a box) were aligned according to the average (center-of-mass) points. (a) The original data segments. (b) A reconstruction of the segment from the 10 angles.

Figure 3, the knee value, approximately 7, remained consistent over the two feature extraction methods.

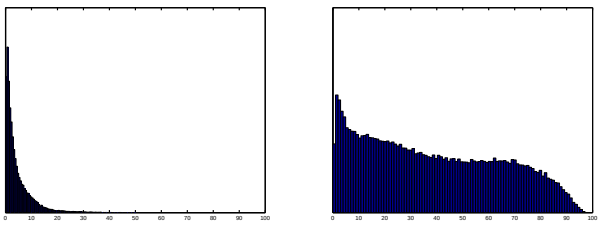
After observing a set of segments which were assigned to states and gathered into clusters, one can easily notice the resemblance and the difference between them; Figure 4 presents a single cluster (one of the seven clusters) obtained as a results of clustering algorithm. Every sub plot, depicts a state derived from the MoG model. The segments are not all alike though most of them have the same structure as can be expected when dealing with a Gaussian distribution. However, there are some prominent differences between them, such as position and size. These two properties were not included as features throughout the process of grouping the segments into states. As a result these two traits could distinguish between different segments in a given cluster. A small segment and a large segment may share one state; the question is not the segment size, but rather, is there a resemblance between their sequences of features (angles).

An empirical validation for our new defined clusters is the kinship between the segments the belong to the same cluster. To observe the resemblance between a cluster-sharing segments as opposed to the difference between none cluster related segments, we computed the histograms of the distances. Figure 5a shows a histogram of the square dif-

ferences between all the segments from the same cluster, and Figure 5b shows a histogram of square differences between segments from different clusters. The results confirm our hypothesis; movement can be segmented into meaningful clusters of movement segments. This segmentation is based on the segments' formal features, on one hand, and on probabilistic considerations on the other. Note that the state-clustering was based on just the markovian relation between states without any explicit information about the trajectory model associated with each state. In spite of that, as can be observed from both Figure 4 and Figure 5, the movement-trajectories related to states within a cluster are similar.

6. CONCLUSION

In this study we analyzed hand-movements. We have first utilized a parametric representation for each data-fragment which was then used to group the segments according to the shape. Next we applied the proposed clustering method to form clusters of trajectories. This last step can be viewed as a shift from a "phone" based representation towards a more high-level "phoneme" based representation. The entire model provides a generative description of the monkey recorded drawings. On going work is focused on the question whether these drawing primitives correspond to different states of brain activity. The current results encourage us to implement this fragmentation process over the corresponding brain activity. A successful match between basic movement fragments and basic brain activity fragments may support the notion that each basic movement fragment has its own corresponding brain activity fragment Apart from the application discussed in this work, approximations of Markov-chains with large number of states arise frequently in important applications such as text and speech analysis. Another application is webpage analysis. The Markov chain interpretation of a surfer's behavior, that is usually utilized for webpage ranking, can be also used for clustering the webpages.



(a) (b)

Fig. 5. (a) A squared differences histogram between segments within one cluster of Markov states (b) A squared differences histogram between segments from different clusters of states.

The authors would like to thank Dr. Rotem Drori for providing the data of the monkey's scribling. This work was supported in part by Deutsch-Israelische Projectkooperation (DIP).

7. REFERENCES

[1] N. E. Berthier. Learning to reach: a mathematical model. *Dev. Psychol.*, 1996.

[2] E. Bizzi, S. Giszter, and F. A. Mussa-Ivaldi. Computations underlying the execution of movement: a novel biological perspective. *Science*, 1991.

[3] E. Bizzi, M. C. Tresch, P. Saltiel, and A. dAvella. New perspectives on spinal motor systems. *Nat. Rev. Neurosci.*, 2000.

[4] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. *International Conference on Knowledge Discovery and Data Mining(KDD)*, 2003.

[5] T. Flash and E. A. Henis. Arm trajectory modification during reaching towards visual targets. *J. Cogn. Neurosci.*, 1991.

[6] T. Flash and B. Hochner. Primitives in vertebrates and invertebrates. *Curr. Opin. Neurobiol.*, 2005.

[7] T. Flash and N. Hogan. Coordination of arm movements: An experimentally confirmed mathematical model. *J. of Neuroscience*, 1985.

[8] X. Ge, S. Parise, and P. Smyth. Clustering markov states into equivalence classes using svd and heuristic search algorithms. *AISTATS*, 2003.

[9] C. Hofsten. Structuring of early reaching movements: a longitudinal study. *J. Mot. Behav.*, 1991.

[10] J. Konczak, M. Borutta, H. Topka, and J. Dichgans. The development of goal-directed reaching in infants: hand trajectory formation and joint force control. *Exp. Brain Res.*, 1995.

[11] H. I. Krebs, M. L. Aisen, B. T. Volpe, and N. Hogan. Quantization of continuous arm movements in humans with brain injury. *Proc. Natl. Acad. Sci.*, 1999.

[12] M. J. Mataric. Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics. imitation in animals and artifacts. *The MIT Press*, 2001.

[13] M. Meila and J. Shi. A random walks view of spectral segmentation. *AISTATS*, 2001.

[14] F. A. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Philos Trans R Soc Lond B Biol Sci*, 2000.

[15] N. Slonim and Y. Weiss. Maximum likelihood and the information bottleneck, 2003.

[16] C. C. Tappert. Cursive script recognition by elastic matching. *IBM J. Research and Development, vol. 26*, 1982.

[17] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the Annual Allerton Conference on Communication, Control and Computing*, 1999.

[18] P. Viviani and R. Schneider. A developmental study of the relationship between geometry and kinematics in drawing movements. *Journal of Experimental Psychology*, 1991.