Learning Activity Patterns Performed With Emotion

Qi Wang
Ecole Central Marseille
CMI, 39 rue F. Joliot Curie,
F-13453,Marseille,France
qi.wang@centrale-marseille.fr

Thierry Artières
Ecole Centrale Marseille
CMI, 39 rue F. Joliot Curie,
F-13453,Marseille,France
thierry.artieres@centrale-marseille.fr

Yu Ding
Telecom ParisTech
LTCI,46 Rue Barrault,
F-75634,Paris,France
yu.ding@telecom-paristech.fr

ABSTRACT
This paper is a preliminary work towards the design of a model able to generate realistic motion sequences conditioned on a number of contextual variables like age, morphology, emotion etc. We focus in a first step on the design of contextual markovian models able to perform recognition of activities performed under various emotions even in the case no training samples are available for a particular (activity, emotion) pair, a zero shot learning setting.

Author Keywords
activity recognition; activity with emotion; virtual agent;machine learning

ACM Classification Keywords
I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism-Animation;

INTRODUCTION
There has been a number of works that aim at designing systems able to synthesize realistic animations of avatars using statistical models. Of course the way one moves depends on many factors which are related to rather stable features such as her morphology age, habits, emotion and on more contextual features such as emotion, role etc. We are interested in designing a generic synthesis system able to generate realistic trajectories in various contexts related to the gender, the age, the morphology, the emotion, the role... of the character to animate. Moreover we would like our generic system to be easily controled by setting these context variables (the gender, the age, the morphology...) by hand before generating an animation. Actually it is quite likely that not all combinations of these contextual variables will occur in the training set or that some combinations will be rare. To overcome these problems and thus to allow learning of statistical models from limited training datasets, one has to design methods able to disentangle the influence on these many contextual factors on the final animation. This latter property could make possible the generation of synthesis whatever the combination of the contextual variables, would this combination occur or not in the training set. This work is a preliminary step in this direction.

HMMs for learning information from motion capture data have been studied in the past decade[2, 6, 11]. In [2], an entropy-based model is presented for learning stylistic vector for each set of activity sequences, which can easily change the styles of an activity. Other data-driven models are studied for dealing with style learning of human motions[10, 1, 9]. Although researchers have achieved some great results, learning style representation is still a challenging task. Besides, [12] proposed a supervised model called parametric HMM which is able to find the linear contextual information in test set by given the prior knowledge in training step. In [8], a contextual HMM framework (CHMMs) is proposed for conditioning the probability distribution of a HMM. These pioneer works showed the ability of Contextual HMMs to perform high quality avatar animation from speech [3] and for laughter animation [4]. researchers have applied parametric HMMs in sythesis of movements of human body, such as [6]. In these works, contextual parameters are given in training step, but more often representing motion style is difficult. So, our work focuses on learning a represention for emotion or any other contexts by taking advantage of CHMMs framework.

We follow this latter line of work and we propose a CHMM approach for learning motion patterns performed in various contexts without any prior knowledge about this contextual information. We extend previous works to jointly learn how to exploit the contextual information and how to represent it. We apply our framework on activity recognition task performed under various emotion. This is a first step towards synthesis which is relevant since first it may be evaluated with objective measure such as recognition accuracy and second since it has been observed in previous studies that the models that allow high recognition accuracy are the ones that yield high quality synthesized trajectories [3].

PROPOSED METHOD
Contextual HMMs (CHMMs) [8] have been originally proposed as an extension of parametric HMMs [12] for dealing with variability in the data. The main idea is to make HMMs parameters (Gaussian means, Covariance matrices and transition probabilities) dependent on what is called external variables that might represent additional contextual variables like the gender of a gesturer in a gesture recognition task, the estimated signal to noise ratio in a speech recognition task...

MOCO'16, July 05-07, 2016, Thessaloniki, GA, Greece
2016 ACM. ISBN 978-1-4503-4307-7/16/07...$15.00
DOI: http://dx.doi.org/10.1145/2948910.2948958

The contextual variables then represent some available additional information that bears some information on the data which is not easy to integrate in a HMM based recognition system. Such a modeling framework has been used both to design accurate recognition systems and to design synthesis system that are conditioned on some input, e.g. to animate an avatar based on its speech signal [3].

In a CHMM, Gaussian means are parameterized. The mean of a Gaussian distribution in state j, $\hat{\mu_j}$, is defined as:

$$\hat{\mu}_j = \bar{\mu}_j + W_j\theta \tag{1}$$

where $\bar{\mu}_j \in \mathbb{R}^d$ is an offset vector and $W_j \in \mathbb{R}^{d\times c}$ are parameters that determine how θ modifies the offset mean $\bar{\mu}_j$ (d is the dimension of the observation space and c the dimension of θ, $c = |\theta|$, i.e. number of contextual variables). If θ variables were useless and did not include information to model the data, then W_j could be set to 0 and one would recover a standard HMM whose Gaussian distribution in state j is $\bar{\mu}_j$. Contextual HMMs are learned with a specific instance of the Baum-Welch EM algorithm [7], [8]. The EM maximization step has an analytical solution which leads to the parameter reestimation formula below which are iterated, as well as the standard parameter reestimation formula for other parameters, until convergence:

$$W_j = \left[\sum_{k,t} \gamma_{ktj}(x_{kt} - \bar{\mu}_j)\theta_k^T\right]\left[\sum_{k,t} \gamma_{ktj}\theta_k\theta_k^T\right]^{-1} \tag{2}$$

where γ_{ktj} denotes the probability of being in state j at time t, given the observed sequence k, θ_k stands for the contextual variables associated to sequence k, and x_{kt} denotes the observation vector (the frame) at time t in sequence k. Note that when θ is known, a CHMM model, that we note λ hereafter, may be instanciated as a standard HMM by computing the means of Gaussian distributions according to Eq. 1. Then any standard algorithm like Viterbi may be used with this model. We note λ_θ the HMM which is the instance of a CHMM model λ obtained given θ value.

We will focus here on rather simple CHMMs where only the means of Gaussian distributions are parameterized by θ. Transition probabilities and covariance matrices are unparameterized, hence independent of θ. Note also that while θ might vary with time we only deal here with a unique θ (hence time-invariant) per sequence.

Learning Activity Models under Different Contexts

We investigate the use of CHMMs for efficiently learning models of activity while samples of these activities have been performed in a few contexts, e.g. various emotions.

Assume that there are M activities $\mathbb{A} = \{A_1, ..., A_M\}$ performed under N different contexts $\mathbb{C} = \{C_1, ..., C_N\}$ (e.g. emotions) in the training set. Any training sequence then comes with a pair of labels (a, c), namely its corresponding activity a and emotion c. There are two usual ways for dealing with this situation. The first one is to learn one model per activity A_j from all training sequences corresponding to this activity, i.e. training sequence x with label (A_j, c). This model, being learned for all the contexts, should be designed to be robust to variability brought by the various contexts, which is not easy to achieve. Alternatively one can learn $M \times N$ models, i.e. a model for every possibility of activity and context (A_i, C_j). The problem lies here in that the training set for a specific pair (A_i, C_j) might be too small (even empty!) to correctly learn all of these models

Using CHMMs offers an alternative. One can indeed learn one CHMM per activity A_j, noted λ^j, from all training sequence x whose label (a, c) is such that $a = A_j$, but using the contextual variables θ to encode the context c of x, $c(x)$. Doing so one can smartly handle variability brought by the contexts, while sharing a number of parameters (transition probabilities, covariance matrices and mean offsets $\bar{\mu}$) enabling learning from few training samples for each (A_i, C_j) pair. One key issue is to define the encoding of the contexts $\{\theta(c), c \in C\}$ since these will strongly affect the behavior of the method. A simple and natural choice is to use what is known as one-hot-codes (we investigate smarter schema in the next section) : The vector $\theta(C_j)$ corresponding to the j^{th} context C_j is a M dimensional vector whose components are all set to zero except the j^{th} one which is set to one. At test time, the context is unknown and activity recognition for a sequence x is performed according to :

$$\hat{a} = argmax_i \left[argmax_j p(x|\lambda^i_{\theta(C_j)})\right] \tag{3}$$

Joint Learning of CHMMs and Context's Representations

An interesting setting that we want to consider is to learn both the parameters of CHMMs models and the representation $\theta(C_j)$'s of all possible contexts. Actually the use of one-hot-codes as above is limited. A more interesting encoding of contexts would be a low dimensional and dense (as opposite to the sparse one-hot-codes) vector representation. Such an encoding would allow more parameter sharing between contexts, thus improving recognition ability of the system for rarely observed (activity, context) pairs in the training set. Further this increased sharing between all parameters might hopefully allow to generalize on unobserved (activity, context) pairs in the training set, as we will demonstrate. We focus now on the case where instead of being given the θ's for all contexts at training time , we don't know what these θ are or should be and we want to learn them together with W_j's and with all other parameters of the CHMMs. Unfortunately since θ's and W_j's interplay there is no closed form solution to embed in a standard EM algorithm. Instead we propose a coordinate ascent like algorithm that alternates between optimizing the CHMM parameters (W's and HMM parameters) while θ's are kept fixed, and optimizing θ's while CHMM parameters remain fixed (See Algorithm 1). In that case both steps may be efficiently performed with standard EM. Reestimation formulas for θ's may be obtained analogously to equation 2, given that all other parameters are kept fixed. Naturally, updating θ_l is based on all training sequences x whose context label c is C_l. The sums above range over all sequences k whose activity label is l, i.e. $c(x^k) = l$.

$$\theta_l = \left[\sum_{k,t,j} \gamma_{ktj} W_{kj}^T \Sigma_{kj}^{-1} W_{kj}\right]^{-1}\left[\sum_{k,t,j} \gamma_{ktj} W_{kj}^T \Sigma_{kj}^{-1}(x_{kt} - \bar{\mu}_{kj})\right] \tag{4}$$

Algorithm 1 Joint training of CHMMs and of contextual vectors

procedure TRAINING
  Train one HMM per activity, yielding M HMMs.
  Initialize CHMMs from HMMs
  Randomly initialize θ vectors
  repeat
    Fix θ's and optimize W's using EM (Cf. Eq. 2)
    Fix W's and optimize θs using EM (Cf. Eq. 4)
  until Convergence
end procedure

EXPERIMENTS AND RESULTS

We performed experiments on activity classification where activities are performed under different emotions. All along the experimental section the contexts C_j stand for the emotion under which an actor performs an activity.

Motion Capture DataSet

We used the dataset in [5]. It consists of motion capture data gathered when 12 actors were performing 8 different activities such as 'Being Seated', 'Sitting Down', 'Knocking on the Door', 'Simple Walk', 'Walk with something in the Hands'... Every activity is performed in 8 different emotional contexts which are 'Anger', 'Anxiety', 'Joy', 'Neutral', 'Panic Fear', 'Pride', 'Sadness' and 'Shame'. Each of the 12 actors performs 12 times each activity-emotion combination. There is then a total of 9216 sequences in the dataset. Motion sequences are represented as sequences of frames (of length between $200-500$ depending on the activity) where each frame is a 69-dimensional vector whose components correspond to 69 joint angles of 23 joints in skeleton. We use PCA to reduce the number of the dimensions of body pose to 20. We report experimental results on only two similar activities: "Simple Walk"(SW) and "Walk with something in the Hands" (WH) which are much similar and confusable.

Experimental Setting

We report averaged experimental results gained with cross validation on few splits of the training, validation and test data. Stratified cross validation is used so that data from all actors occur in the three datasets. Also in all these experiments we select a few (activity, emotion) pairs that we make only appear in the test set (we call these *missing pairs*) to investigate if the models that we compare may generalize well to cases that have not been observed in the training stage. Test set performances are then reported as *Test performance* based on the results on the test set except on *missing pairs* and as *Missing pairs performance* computed on *missing pairs* only. We compare our approach with a HMM baseline system where we use one HMM per (activity, emotion) combination. In any case we use one CHMM per activity. Whatever the models (HMMs, CHMMs) each model is a ergodic model (i.e. all transitions are allowed) with 12 states and we use single Gaussian emission probability densities. All models are trained through Maximum Likelihood Estimation. We trained standard HMM first to initialize most of CHMM parameters, initial state probabilities $\{\pi_i\}$, transition probabilities $\{a_{ij}\}$, means of Gaussian emission probability densities $\{\bar{\mu}_j\}$, and co-variance matrices of Gaussian emission probability densities $\{\Sigma_j\}$. Once HMMs are trained, CHMMs are built by copying the HMM parameters and by initializing W's parameters to small random values, then the CHMM parameters are refined using formulas as in Eq. 2 and/or Eq. 4.

Activity Classification

We first report activity classification results (Figure 1). We investigated two settings. In any case recognition of a test sample is performed without any information on it (i.e. the corresponding emotion is unknown). We report results in two settings, by learning with all training data available (models are named *CHMM+* and *HMM+*) and learning with small training sets by randomly choosing one fourth of available training material (models are named *CHMM-* and *HMM-*).

The most advantageous experimental setting here is when one has at his disposal a complete training set, i.e. training samples are available for any label pair (activity, emotion). This setting corresponds to the curves *HMM Train*, *HMM Valid* and *HMM Test* (resp. *CHMM Train/validation/Test*) which show the performance of the HMM baseline and of our CHMM approach on the three datasets as a function of the dimension of θ (HMM performances are independent θ and plotted as constant). One may see that the while CHMM is outperformed by the HMM baseline on all datasets when enough training data is available ($+$ curves), CHMM outperform HMMs when the training data set is smaller ($-$ curves). This comes from the fact that there is a parameter sharing schema between activity-emotion models in CHMMs. Also it is worth noticing that the dimension of θ does not seem to impact much the performance here, meaning one may significantly compress the size of the activity models with respect to the baseline HMM system without decreasing accuracy.

The figure also shows four additional curves with are called *HMM*$+/-$ *Missing Pairs* and *CHMM*$+/-$ *Missing Pairs*. These curves correspond to the performance on *Missing Pairs* data only, a more difficult setting. Both our approach and the HMM baseline naturally perform lower on these data but CHMM performance drop is significantly smaller than the one of HMMs, especially when less training data is available.

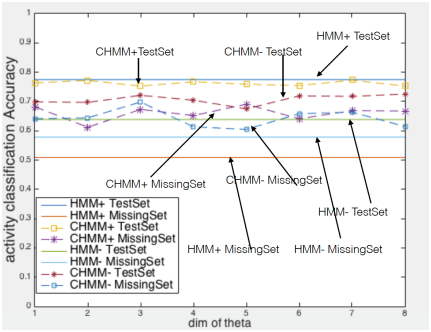

Figure 1. Accuracy of Activity Classification wrt. $|\theta|$s.

A Deeper Analysis

We provide some addition experimental results that provide some light what is happening. Figure 2 shows the trajectories described by the θ vectors of the eight emotions along

the training process. One sees for instance that the representations of *anxiety* and *shame* move away from one another while *anger* and *joy* become closer. As these results show emotion is clearly taken into account by the markovian models and actually help improve modeling accuracy. Yet despite these encouraging results emotion recognition results, as shown in Figure 3 and Fig.4, are rather disappointing. Here the task is to recognize emotion when activity is supposed to be known at test time. One sees that the accuracy is reasonably good for the training, validation and test sets when the dimension of θ increases and also that as before CHMMs outperform HMMs in small training set experiments while HMMs are more accurate with large training sets. Looking at missing pairs performance (HMMs cannot perform on these data) it seems quite puzzling that the performance drops to almost random when computed on missing pairs only. This seems a little contradictory with previous promising results and we don't have clear explanations for this phenomenon.

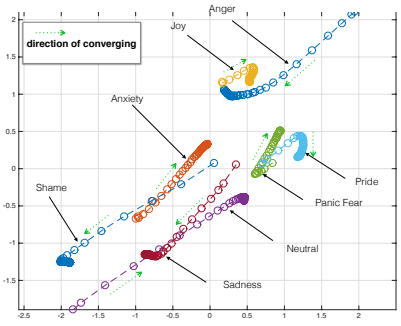

Figure 2. Trajectories of the 2D representations (θ) of the 8 emotions along the training process (random initialization).

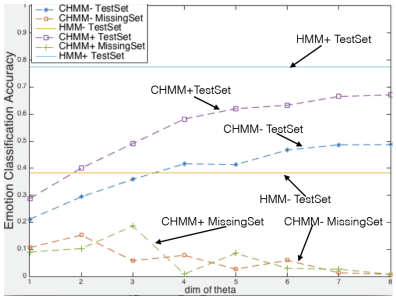

Figure 3. Accuracy of Emotion Classification wrt. $|\theta|$.

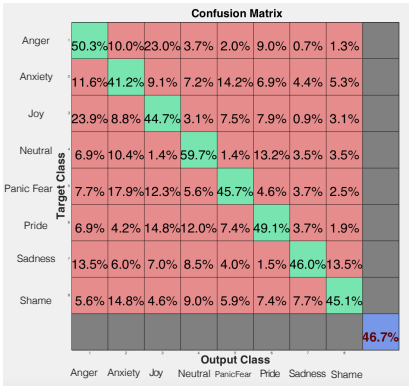

Figure 4. Confusion Matrix of CHMM- on Test Set $|\theta| = 6$

CONCLUSION

We described an approach for activity recognition performed under various contexts. We showed how this may be handled with contextual markovian models by learning simultaneously activity models' parameters and representation of the contexts. Preliminary experiments show that our proposal enables recognizing an activity performed under an emotion for which there was no training samples.

ACKNOWLEDGMENTS

We thank Catherine Pelachaud (LTCI, Telecom ParisTech, France) for her help and advice on conducting this work. Qi WANG's thesis is funded by the China Scholarship Council.

REFERENCES

1. Alemi, O., Li, W., and Pasquier, P. Affect-expressive movement generation with factored conditional restricted boltzmann machines. In *Affective Computing and Intelligent Interaction, International Conference on*, IEEE (2015), 442–448.
2. Brand, M., and Hertzmann, A. Style machines. In *Proc.27th Conf. Computer graphics and interactive techniques* (2000).
3. Ding, Y., Pelachaud, C., and Artières, T. Modeling multimodal behaviors from speech prosody. In *IVA* (2013).
4. Ding, Y., Prepin, K., Huang, J., Pelachaud, C., and Artières, T. Laughter animation synthesis. In *AAMAS* (2014).
5. Fourati, N., and Pelachaud, C. Multi-level classification of emotional body expression. In *FG* (2015).
6. Herzog, D., and Krüger, V. Recognition and synthesis of human movements by parametric hmms. In *Statistical and Geometrical Approaches to Visual Motion Analysis*. Springer, 2009, 148–168.
7. Rabiner, L. R., and Juang, B.-H. An introduction to hidden markov models. *ASSP Magazine, IEEE 3*, 1 (1986), 4–16.
8. Radenen, M., and Artieres, T. Contextual hidden markov models. In *ICASSP* (2012).
9. Taubert, N., Christensen, A., Endres, D., and Giese, M. A. Online simulation of emotional interactive behaviors with hierarchical gaussian process dynamical models. In *Proc. the ACM Symposium on Applied Perception*, ACM (2012), 25–32.
10. Taylor, G. W., and Hinton, G. E. Factored conditional restricted boltzmann machines for modeling motion style. In *Proc.the 26th ICML*, ACM (2009), 1025–1032.
11. Tilmanne, J., Moinet, A., and Dutoit, T. Stylistic gait synthesis based on hidden markov models. *EURASIP 2012*, 1 (2012), 1–14.
12. Wilson, A. D., and Bobick, A. F. Hidden markov models for modeling and recognizing gesture under variation. *International Journal of Pattern Recognition and Artificial Intelligence 15*, 01 (2001), 123–160.