Technical Gestures Recognition by Set-Valued Hidden Markov Models with Prior Knowledge

Yann Soullard[a], Alessandro Antonucci[b], Sébastien Destercke[a]

Abstract Hidden Markov models are popular tools for gesture recognition. Once the generative processes of gestures have been identified, an observation sequence is usually classified as the gesture having the highest likelihood, thus ignoring possible prior information. In this paper, we consider two potential improvements of such methods: the inclusion of prior information, and the possibility of considering convex sets of probabilities (in the likelihoods and the prior) to infer imprecise, but more reliable, predictions when information is insufficient. We apply the proposed approach to technical gestures, typically characterized by severe class imbalance. By modelling such imbalances as a prior information, we achieve more accurate results, while the imprecise quantification is shown to produce more reliable estimates.

1 Introduction

In this paper we are concerned with classification tasks where one wants to identify gestures (a popular computer vision task [4]) as well as errors in incorrectly executed gestures. We assume the possible gestures belong to a set $\mathcal{C} := \{c_1, \ldots, c_M\}$ and denote as C the variable taking values in $\mathcal{C}$. A gesture recognition algorithm then aims at assigning the correct value $c^* \in \mathcal{C}$ to a given sequence. With few exceptions [5], gestures are regarded as multivariate time series, say $(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T)$, with $\boldsymbol{o}_t \in \mathbb{R}^F$ the joint observation of the F features extracted from the t-th frame, for each $t = 1, \ldots, T$. Technical gestures are quite specific, as they are based on particular movements, they require specific skills and they should be executed with a high level of precision. Examples of technical gestures can be found in many domains such

E-mail: {yann.soullard,sebastien.destercke}@hds.utc.fr, alessandro@idsia.ch.
[a]Heudiasyc, University of Compiègne, Compiègne (France).
[b]Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Manno-Lugano (Switzerland).

as sport (e.g., the forehand of a tennis player), manufacturing (e.g., doing a welding), or handicraft (e.g., the movements of a potter), just to cite a few.

Technical gestures are confronted with specific problems. First, due to the fact that most learning data have to be collected from experts (e.g., if in a later employee training stage, we want to recognize well and badly performed gestures), the obtained data sets are typically small and imbalanced. Those data can also be quite noisy, as measurements are often performed in working environments. Also, when the recognition model is used to decide if a task or a gesture has been performed correctly, a recognition error might have a significant economic impact (e.g. the manufacturing of a defective part or an interruption in the production line). This is why considering tools able to account for this imbalance or this lack of data is important.

Hidden Markov Models (HMMs, [9]) are probabilistic graphical models that can easily cope with multivariate time series, and are therefore often used for gesture recognition [2, 6]. As they are generative models usually trained with maximum-likelihood estimates, HMMs are less prone to over-fitting than their discriminative counterparts [10]. However, they can still suffer from bad parameters estimation when the training examples do not fit well the true data distribution [3]. To gain reliability in the learning, a recent paper [1] proposed a set-valued quantification of the HMM parameters inspired by the theory of imprecise probabilities, for which polynomial-time inference algorithms have been also developed [7]. With those imprecise HMMs, evidential information might not be sufficient to unequivocally recognize the performed gesture, and sets of candidate gestures might be obtained instead. Sect. 2 contains background information about imprecise methods and HMMs.

Such approaches take care of the limited amount of available data, while the imbalances over the classes (a typical issue for data of this kind) are neglected by implicitly assuming a uniform marginal distribution over the gestures. The main methodological contribution of this paper, explained in Sect. 3, is a procedure to add prior information about the classes, that can itself be imprecise and represented as a convex set of probability mass functions. The methodology is validated in Sect. 4 on technical gestures performed in an aluminum foundry. This real-world application is part of a training system in a virtual environment for tasks related to mold cleaning (Fig. 1).

2 Background

Imprecise Probability. Let C denote the class variable associated to the gesture and $\mathcal{C}$ the M possible values. If the uncertainty about C is described by a probability mass function P, the task of deciding the actual value of C, assuming zero/one losses, returns:

$$c_P^* := \arg\max_{c\in\mathcal{C}} P(c)\,. \tag{1}$$

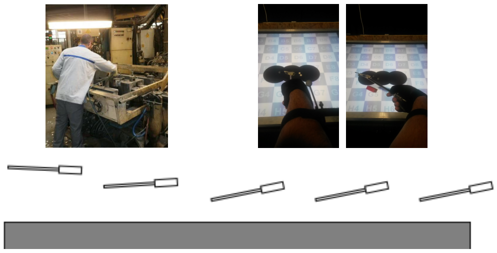

Fig. 1 Pictures of mold cleaning in a work environment (top left) and in the experimental station of a virtual environment (top right). Expected positions and inclinations of a blower during a technical gesture with a movement from the right to the left (bottom).

In many cases single probability mass functions might be unable to provide a reliable uncertainty model. Assume for instance that, among three possible gestures, an expert is telling us that c_1 is at least as probable as c_2, which is in turn at least as probable as c_3. Deciding that $P(C) = [.7, .2, .1]$ is a better model than $P'(C) = [.6, .3, .1]$ from this information alone is questionable. In such situations, *credal sets*, i.e., closed convex sets of probability mass functions, can offer a more cautious, hence reliable, uncertainty model. In our case, a credal set over $\mathcal{C}$, denoted $K(C)$, will be specified by a finite number of linear constraints, or equivalently by its (finite) set of extreme points. In the expert example with three gestures, we can consider the credal set $K(C)$ defined by the constraints $P(c_1) \geq P(c_2) \geq P(c_3)$, together with non-negativity and normalization, or equivalently, by listing the extreme points $P_1(C) = [1, 0, 0]$, $P_2(C) = [\frac{1}{2}, \frac{1}{2}, 0]$, and $P_3(C) = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ (Fig. 2). The generalization of Eq. (1) to credal sets can be achieved in many ways. Here we consider the *maximality* criterion, which returns the following sets of optimal classes:

$$\mathcal{C}^*_K := \{c' \in \mathcal{C} : \nexists\, c'' \in \mathcal{C} \,\text{s.t.}\; P(c'') > P(c') \,\forall\, P(C) \in K(C)\}\,. \tag{2}$$

Non-optimal classes are therefore those such that, for each element of the credal set, there is another class with strictly higher probability.

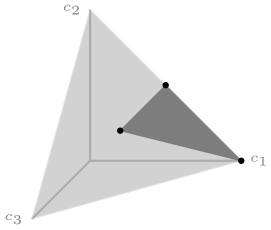

Fig. 2 A credal set modeling uncertainty about a gesture with three options.

Hidden Markov Models (HMMs). HMMs [9] are popular probabilistic descriptions of time series with many applications in speech recognition and computer vision, to name but a few. HMMs assume the observation $\boldsymbol{O}_t$ is generated by a paired *state* variable X_t, for each $t = 1, \ldots, T$, with T the length of the sequence. State variables are in turn assumed to be generated by a Markov chain process. All state variables take their values from a space $\mathcal{X}$ of cardinality N. An HMM specification comprises an initial state probability mass function $P(X_1)$, a $N \times N$ state transition probability matrix $P(X_{t+1}|X_t)$, and a (usually normal) distribution for each observation with mean and covariance indexed by the corresponding state, say $\boldsymbol{\mu}(X_t)$ and $\boldsymbol{\sigma}(X_t)$. We consider *stationary* models with the values of the parameters independent of t. HMMs give a compact specification of the joint density:

$$P(x_1, \ldots, x_T, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_T) := P(x_1) \prod_{t=1}^{T-1} P(x_{t+1}|x_t) \prod_{t=1}^{T} \mathcal{N}_{\boldsymbol{\sigma}(x_t)}^{\boldsymbol{\mu}(x_t)}(\boldsymbol{o}_t) \,. \tag{3}$$

By marginalizing the states in Eq. (3) we obtain the *likelihood* of a sequence $P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T)$. This can be achieved in $O(TN^2)$ time by a message propagation algorithm [9]. HMMs are trained using an Expectation-Maximization approach, the Baum-Welch algorithm, detecting a local maximum of the likelihood defined by the joint probabilities of the training sequences and of their classes. Classification can then be achieved by: (i) training a HMM per class; and then (ii) assigning to a test sequence $(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T)$ the class associated to the HMM giving the highest likelihood to the sequence, i.e.,

$$c^* := \arg\max_{c \in \mathcal{C}} P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c) \,, \tag{4}$$

where notation $P(\ldots|c)$ is used for the density corresponding to the HMM associated to class c. Here no prior probabilities over the classes are supposed to be available, i.e., a uniform distribution over them is implicitly assumed.

As Baum-Welch estimates might be unreliable, for instance when using few data or short sequences, imprecise probabilities have been proposed to mitigate this unreliability in the HMM quantification [1]. An HMM with imprecise parameters can be learned from a sequence by combining the Baum-Welch algorithm with the *imprecise Dirichlet model* (IDM, [11]). In this model, $P(X_1)$ is replaced by a credal set $K(X_1)$ and $P(X_{t+1}|x_t)$ with $K(X_{t+1}|x_t)$ for each x_t. As shown in [7], the bounds $[\underline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c), \overline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c)]$ of the likelihood with respect to those credal sets can be computed with the same time complexity as the precise computation. The classification scheme in Eq. (4) can then be extended to set-valued HMMs by comparing the likelihood intervals and then deciding the optimal ones as in Eq. (2).

3 HMM-based Classification with Prior Knowledge

If prior knowledge about the classes is available in the form of a mass function $P(C)$, the likelihood-based classification scheme in Eq. (4) becomes:

$$c^* = \arg\max_{c \in \mathcal{C}} P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T | c) \cdot P(c)\,, \tag{5}$$

which corresponds to a comparison of the posterior probabilities

$$P(c|\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T) \propto P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T | c) \cdot P(c)\,. \tag{6}$$

A proper assessment of the prior mass function is clearly crucial in this Bayesian framework. Yet, the elicitation of qualitative or quantitative expert prior knowledge suffers from the same issues discussed in Sect. 2, and a credal set $K(C)$ might offer a more reliable model of the prior knowledge about C. We therefore consider a twofold generalization of Eq. (5) to imprecise probabilities in which $P(C)$ is replaced by a credal set $K(C)$, and the sequence likelihoods $P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c)$ are replaced by their lower/upper bounds learned from the training data. The optimal classes can be therefore obtained by applying the criterion in Eq. (2) to the, imprecisely specified, posterior probabilities in Eq. (6). To achieve that in practice, given two classes $c', c'' \in \mathcal{C}$, we evaluate whether the posterior probability for c'' is always greater than that of c', i.e.,

$$\min_{\substack{P(C) \in K(C) \\ P(\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T|C) \in [\underline{P}(\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T|C), \overline{P}(\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T|C)]}} \frac{P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T | c'') \cdot P(c'')}{P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T | c') \cdot P(c')} > 1\,. \tag{7}$$

where we assume the denominator strictly positive. If the above inequality is satisfied, class c' is removed from the set of optimal labels. The set of optimal options $\mathcal{C}^*_K$ is obtained by iterating the test in Eq. (7) for any pair of classes, and removing from $\mathcal{C}$ the dominated options. The optimization with respect to the imprecisely specified likelihoods is trivial and allows to rewrite Eq. (7) as follows:

$$\min_{P(C) \in K(C)} \frac{\underline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T | c'') \cdot P(c'')}{\overline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T | c') \cdot P(c')} > 1\,. \tag{8}$$

As $K(C)$ can be expressed by linear constraints, the task in Eq. (8) is a linear-fractional task, which can be reduced to a linear program and solved in polynomial time w.r.t. the number of classes M by a linear solver.

4 Empirical Validation

We test the proposed approach on six technical gesture data sets (Tab. 1). The TG and TGE datasets refer respectively to classification of types of gestures and types of errors (for specific gestures). The gestures are performed in an aluminium foundry and refer to a workstation where a technician cleans a mold (Fig. 1). The technician performs several tasks with different tools such as a compressed-air blower, a scraper and a pistol. Motion capture is performed by markers attached to the tools and the user's body. Markers are tracked by infrared cameras and, at each time frame, 3D positions and orientations are extracted. Such raw features may not directly provide a good modelling of the gesture. Following [8], we compute high-level features such as velocities, pairwise distances and angles to enrich the description.

Dataset	F	M	Samples for $c_1/\dots/c_m$
TG_1	15	4	320/160/224/287
TG_2	15	4	192/320/256/287
TG_3	18	4	100/100/40/20
TGE_1	19	4	57/36/45/33
TGE_2	4	3	15/30/20
TGE_3	4	3	20/10/15

Table 1 Number of features, classes, and samples per class in the benchmark.

To train HMMs as in Eq. (3), we run the Baum-Welch algorithm with a maximum of 25 iterations before convergence and three states for the hidden variables (i.e., $N = 3$). For the imprecise quantification we set $s = 4$ for the parameter determining the imprecision level (in term of missing observations) in the IDM. The *accuracy* (i.e., the percentage of properly classified gestures) describes the performance of the precise classifiers. We say that an imprecise classifier is *indeterminate* when more than one class is returned as output. To characterize the output of an imprecise classifier we use its *determinacy* (i.e., percentage of determinate outputs) and *output size* (i.e., average number of classes in output when indeterminate). The performance is described in terms of *single accuracy* (i.e., accuracy when the output is determinate) and *set accuracy* (i.e., percentage of indeterminate outputs including the true class). For a direct comparison with precise classifiers we compare the accuracy with the u_{80} utility-based measure. This is basically a positive correction (namely $1.2(q-1)/q^2$), advocated in [12], of a discounted accuracy giving $1/q$ to a classifier returning q options if one of them is correct, and zero otherwise.

The proposed method is intended to achieve robustness when coping with small datasets. Accordingly, we adopt a (five-fold) cross validation scheme with one fold for training, and the rest for testing. In Fig. 3, we compare the accuracies of the approaches based on the likelihood (Eq. (4)) and the posterior (Eq. (6)) with the u_{80} for the imprecise posterior. The precise prior is obtained from the distribution over the classes of the training data. The prior credal set is similarly obtained by the IDM ($s = 4$). Introducing the prior has a positive effect which is only modest in the precise case and more notable in the imprecise case. A deeper analysis of the imprecise model based on the posterior is in Tab. 2. Remarkably, the classifier achieves high determinacies and, when indeterminate, only two classes are typically returned. The single accuracies are higher than the accuracies of the precise models (i.e., when determinate the imprecise classifier outperforms the precise methods). Finally, on two datasets, when indeterminate the imprecise classifier returns always two classes and one of them is always the correct one.

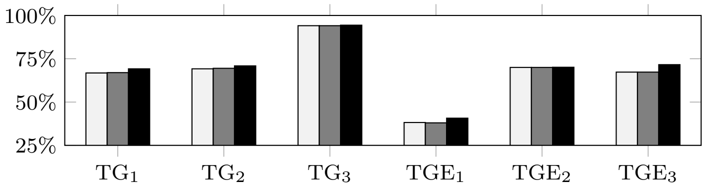

Fig. 3 Accuracies of the likelihood (white) and posterior (gray) comparison against the u_{80} of the imprecise posterior (black).

Dataset	Precise accuracy	Single accuracy	Set accuracy	Determinacy	Output size
TG$_1$	67.3%	70.3%	80.8%	93.0%	2.1
TG$_2$	69.7%	71.5%	78.8%	93.7%	2.1
TG$_3$	94.7%	95.0%	100.0%	97.9%	2.0
TGE$_1$	38.0%	40.7%	58.7%	96.2%	2.0
TGE$_2$	70.0%	71.1%	76.7%	94.6%	2.0
TGE$_3$	67.3%	71.1%	100.0%	93.6%	2.0

Table 2 Performance of the classifier in the precise and imprecise posterior case.

5 Conclusions and Outlooks

A new classification algorithm for multivariate time series is proposed. The sequences are described by HMMs, and the likelihoods returned by these models are combined with a prior distribution over the classes. A robust modeling based on an imprecise-probabilistic quantification of the HMM parameters and the prior is shown to produce more reliable classification performance, without compromising the computational efficiency. Such an approach allows to deal with small and imbalanced datasets. We obtain a set of predicted labels when the information is not sufficient to recognize the performed gesture. An application to technical gesture recognition in an industrial context is reported. As future work, we want to apply our approach to sequences of gestures, by also achieving a segmentation of the various gestures.

Acknowledgments

This work is founded by the European Union and the French region Picardie. Europe acts in Picardie with the *European Regional Development Fund (ERDF)*. This work is supported by the ANR UML-net project, grant ANR-14-CE24-0026 of the French *Agence Nationale de la Recherche (ANR)*.

References

1. A. Antonucci, R. de Rosa, A. Giusti, and F. Cuzzolin. Robust classification of multivariate time series by imprecise hidden Markov models. *International Journal of Approximate Reasoning*, 56(B):249–263, 2015.
2. F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. *Gesture in Embodied Communication and Human-Computer Interaction: 8th International Gesture Workshop, GW 2009, Revised Selected Papers*, chapter Continuous Realtime Gesture Following and Recognition, pages 73–84. Springer, 2010.
3. G. Bouchard and B. Triggs. The tradeoff between generative and discriminative classifiers. In *International Symposium on Computational Statistics*, pages 721–728, 2004.
4. A. Chaudhary, J.L. Raheja, K. Das, and S. Raheja. Intelligent approaches to interact with machines using hand gesture recognition in natural way: a survey. *International Journal of Computer Science and Engineering Survey*, 2(1):122–133, 2011.
5. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2247–2253, 2007.
6. K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz. Multi-HMM classification for hand gesture recognition using two differing modality sensors. In *Circuits and Systems Conference (DCAS)*, pages 1–4. IEEE, 2014.
7. D.D. Mauá, A. Antonucci, and C.P. de Campos. Hidden Markov models with set-valued parameters. *Neurocomputing*, 180:94–107, 2015.
8. N. Neverova, C. Wolf, Taylor G.W., and F. Nebout. Multi-scale deep learning for gesture detection and localization. In *ECCV Workshop on Looking at People*, 2014.
9. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
10. Y. Soullard, M. Saveski, and T. Artières. Joint semi-supervised learning of hidden conditional random fields and hidden Markov models. *Pattern Recognition Letters*, 37:161–171, 2014.
11. P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.
12. M. Zaffalon, G. Corani, and D.D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.