Action Recognition by Imprecise Hidden Markov Models

Alessandro Antonucci[1], **Rocco de Rosa**[2], **and Alessandro Giusti**[2]
[1]IDSIA, "Dalle Molle" Institute for Artificial Intelligence, Lugano, Switzerland
[2]Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Milano, Italy

Abstract— *Hidden Markov models (HMMs) are powerful tools to capture the dynamics of a human action by providing a sufficient level of abstraction to recognise what two video sequences, depicting the same kind of action, have in common. If the sequence is short and hence only few data are available, the EM algorithm, which is generally employed to learn HMMs, might return unreliable estimates. As a possible solution to this problem, a robust version of the EM algorithm, which provides an interval-valued quantification of the HMM probabilities is provided. This takes place in an imprecise-probabilistic framework, where action recognition can be based on the (bounds of the) likelihood assigned by an imprecise HMM to the considered video sequence. Experiments show that this approach is quite effective in discriminating the hard-to-recognise sequences from the easy ones. In practice, either the recognition algorithm returns a set of action labels, which typically includes the right one, either a single answer, which is very likely to be correct, is provided.*

Keywords: Human Action Recognition, Hidden Markov Models, Imprecise Probabilities, Credal Networks

1. Introduction

Recognising human activities from video is a natural application of computer vision. Motions, however, inherently possess a high degree of variability, and are subject to a large number of nuisance factors, such as illumination, background, viewpoint, locality. When several objects/persons move in the field of view occlusion problems also arise, even though the presence of objects in the vicinity can help to disambiguate the activity class. Recently, recognition methods which neglect action dynamics (typically extracting spatio-temporal [1] features from the 3D volume associated with a video) have proven very effective. In other works (*e.g.*, [2]) the dynamical aspects of the motion are considered, but focusing only on a very short fragment of the action. However, encoding the dynamics of videos or image sequences by means of some sort of dynamical model can be useful in situations in which the dynamics is critically discriminative.

Furthermore, the actions of interest have to be temporally segmented from a video sequence: we need to know when an action/activity starts or stops. Actions of sometimes very different lengths have to be encoded in a homogeneous fashion in order to be compared ("time warping"). Dynamical representations are very effective in coping with time warping or action segmentation [3]. In these scenarios, action (or identity) recognition reduces to classifying dynamical models.

Hidden Markov models (HMMs) [4], in particular, have been indeed widely employed in action recognition [3] and gait identification [5]. Each action is seen as a temporal sequence of events, whose ordering is exploited by HMMs to capture both appearance and dynamics of the action. HMM classification can happen by simply evaluating the likelihood of a new sequence with respect to the learnt models, or by learning a new model for the test sequence, measuring its distance from the old models according to some dissimilarity measure, and attributing to it the label of the closest model. As the video data only refer to the observational variables, while, by definition, the hidden variables are directly unobservable, an algorithm to learn from incomplete data is necessary to quantify the HMM corresponding to a video sequence. A typical choice to recursively obtain a (local, and hence approximate) maximum-likelihood estimator is the EM algorithm. Yet, for short sequences, the restricted amount of observational data can make the HMM quantification provided by this algorithm very inaccurate.

In order to achieve more robust and hence reliable estimates when coping with small or incomplete data, the theory of imprecise probability [6] suggests the opportunity of working with sets of probability distributions instead of single "precise" estimators. In particular, when learning from multinomial data, the imprecise Dirichlet model (IDM, [7]) provides interval-valued estimates for the probabilities of the different outcomes. These intervals are obtained following a Bayesian-like approach with a set of Dirichlet priors (instead of a single one) modelling a condition of near-ignorance about the parameters.

In the present paper, in order to gain robustness in the estimates, we apply these ideas to EM-based learning of HMMs. As the EM for HMMs computes the expectations for the number of hidden variables (in a given state or doing a particular transition), it is possible to use the IDM to obtain interval-valued estimates by simply considering these expectations instead of the real counts to be used if complete data are available. In this way an imprecise (-probabilistic version of the) HMM, quantified by interval-valued probabilities, can be obtained from a small data set, while in the limit of large data sets the imprecise model collapses into the precise one, as it can be obtained with the standard EM.

Notably, exactly as a standard HMM can be regarded as a (dynamical) Bayesian net, an imprecise HMM corresponds to a *credal net* [8], for which a number of inference algorithms have been developed. In particular, we can efficiently compute lower and upper bounds for the probability assigned by a (imprecise) HMM to a video sequence (*i.e.*, to a joint state of the observable features extracted from each frame). This allows for a likelihood-based classification, which generalises similar approaches already considered in the precise-probabilistic setting (*e.g.*, [9]). Unlike the precise case, when considering the bounds of the likelihood, a partial overlapping between the highest posterior probability intervals corresponding to a condition of indecision about the class (action label) with highest probability can appear. Thus, the proposed algorithm is a *credal* classifier, which can eventually return more than a single action label for the video sequence (*e.g.*, [10]).

An empirical validation, based on two classical benchmarks for video action recognition, shows that this approach is particularly effective in discriminating the hard-to-classify sequences from the easy ones. Classification is performed both with the precise technique based on the standard EM and with its imprecise counterpart. If the imprecise classifier returns a single class, this, by definition, coincides with that provided by the precise method and it is very likely to be correct, while, when multiple options are returned, the precise method is generally wrong.

Other approaches based on measures more informative than the likelihood can be also considered. With these methods we typically gain robustness, and, in spite of a higher average number of classes in output, this set is very likely to include the correct one. Yet, for these methods we still miss a precise counterparts to be used for this kind of discriminative analysis. Overall, the proposed approach can be profitably used as pre-processing tool for action recognition, which returns a single, generally correct, action label for some sequences, while for the others achieves a substantial restriction in the number of candidate labels, this set in need of further restriction by some other procedure.

The paper is organised as follows: in Sect. 2 we review the necessary background material about HMMs, EM and imprecise probabilities; then, in Sect. 3, we propose our imprecise-probabilistic version of the EM algorithm for HMMs; this approach is applied to HMM-based action recognition in Sect. 4, while an empirical validation is in Sect. 5; conclusions and outlooks are finally in Sect. 6.

2. Background

2.1 Hidden Markov Models (HMMs)

HMMs [4] are very popular dynamic probabilistic models intended to describe temporal sequences when coping with uncertainty. The *hidden* layer of the model is a collection of (directly unobservable) variables, one for each (discrete) time step, modelling the actual state of the system. This is *Markovian*, which means that each configuration is only affected by that at the previous time step.

The *observable* layer corresponds to a second collection of variables, again one for each time step. These are intended to report observable information about the system configuration. The configuration of an observable variable is only affected by the relative hidden variable.

If time t varies in $\{1,\ldots,T\}$ (*i.e.*, we have T discrete time steps), the model is defined over variables $\{(X_t,\mathbf{O}_t)\}_{t=1}^{T}$, where the hidden variable X_t takes values in a finite set $\mathcal{X}$, which is the same at any time ($|\mathcal{X}|=N$); F real-valued features are observed, *i.e.*, the observable variables $\mathbf{O}_t$ are assumed to take values in $\mathbb{R}^F$.

We denote by O_t^f the f-th feature (*i.e.*, coordinate of $\mathbf{O}_t$) of the model. The independence assumptions characterising the model induce in the joint density $P(x_1,\ldots,x_T,\mathbf{o}_1,\ldots,\mathbf{o}_T)$ the following factorisation:

$$P(x_1)\prod_{t=1}^{T-1}P(x_{t+1}|x_t)\prod_{t'=1}^{T}\prod_{f=1}^{F}\mathcal{N}(o_{t'}^f)_{\mu^f(x_{t'}),\sigma^f(x_{t'})}, \quad (1)$$

where $\mathcal{N}(o)_{\mu,\sigma}$ is a Gaussian over o with mean μ and standard deviation σ, $\mu^f(x)$ is the mean of the Gaussian of the f-th feature, when the corresponding hidden variable is in the state $x\in\mathcal{X}$, and similarly for the std. Means, standard deviations, and the hidden-to-hidden transitions are independent of t and the features are real-valued: thus, (1) defines a continuous stationary *hidden Markov model*. We denote by λ a generic HMM specification.

2.2 Learning HMMs by EM algorithm

When complete data about both hidden variables and observable features are available, maximum-likelihood estimators can be used for HMM quantification. Yet, hidden variables are by definition unobservable, and algorithms to learn the model from incomplete data should be considered instead. A typical choice is the *expectation maximisation* (EM, [11]): a recursive estimation, which, after a random initialisation, converges to a local maximum of the likelihood. For HMMs, closed formulae can be obtained for each recursion. *E.g.*, for the probabilities on the hidden layer:

$$P^{(\mathrm{new})}(x_1) = \frac{P^{(\mathrm{old})}(x_1,\mathbf{o}_1,\ldots,\mathbf{o}_T)}{\sum_{x_1\in\mathcal{X}}P^{(\mathrm{old})}(x_1,\mathbf{o}_1,\ldots,\mathbf{o}_T)}, \quad (2)$$

$$P^{(\mathrm{new})}(x_{t+1}|x_t) = \frac{\sum_{t=1}^{T-1}P^{(\mathrm{old})}(x_t,x_{t+1},\mathbf{o}_1,\ldots,\mathbf{o}_T)}{\sum_{t=1}^{T-1}P^{(\mathrm{old})}(x_t,\mathbf{o}_1,\ldots,\mathbf{o}_T)} \quad (3)$$

where the right-hand sides of these equations are efficiently computed by the forward-backward algorithm [4], and they can be regarded as ratios of expected number of hidden variables in a given state or doing a particular transition.

Yet, these estimates are sometimes unreliable and unstable with respect to the initialisation of the parameters, especially when only few data are available.[1]

2.3 Imprecise Probability and Imprecise HMMs

The theory of *imprecise probability* [6] is a generalisation of the classical Bayesian theory of probability, where instead of single probability distributions, sets of distributions are assumed to provide a more robust and realistic model of uncertainty. In particular, uncertainty about the state of a variable X is described by a *credal set* $K(X)$, which is a collection of distributions $P(X)$ over X. Inference over a credal set is intended as the computation of the lower and upper bounds, with respect to the whole set of distributions, of the considered expectation. This problem can be solved by considering only the *extreme* points of the credal set, this transforming an optimisation over a continuous domain into a combinatorial task.

The *imprecise Dirichlet model* (IDM, [7]) can be used to learn a credal set $K(X)$ from data. This set is made of all the distributions $P(X)$ consistent with the linear constraints:

$$\frac{n(X=x)}{N+s} \leq P(X=x) \leq \frac{n(X=x)+s}{N+s}, \qquad (4)$$

for each $x \in \mathcal{X}$, where $n(X=x)$ is the number of records such that $X=x$, N the total amount of data, and the hyperparameter s describes the degree of caution in the inferences.

Credal sets have been adopted to extend Bayesian nets to imprecise probabilities. The result is a class of imprecise probabilistic graphical models called *credal nets* [8]. Thus, as well as a standard HMM can be regarded as a Bayesian net, a HMM where the "precise" prior $P(X_1)$ and transition matrix $P(X_{t+1}|X_t)$ have been replaced by a corresponding collection of credal sets, is an *imprecise HMM*, which corresponds to a particular credal net. Despite the NP-hardness characterising inference in general credal nets, a number of algorithms to efficiently compute exact inferences has been developed for same special cases. For our purposes, it is worth noting that the algorithm in [10] can efficiently compute the bounds of the marginal probability for a set of variables in a tree-shaped credal net.

3. An imprecise-probabilistic EM

As noted in Sect. 2.2, EM estimates when only few data are available can be inaccurate. Thus, on the basis of the discussion in Sect. 2.3, in these cases, it seems quite natural to gain robustness by formulating an imprecise-probabilistic version of the EM algorithm. This becomes particularly simple by exploiting the fact that recursions in (2) and (3) have been already formalised in terms of expected counts. Thus, we can obtain interval-valued estimates by replacing the integer counts required by the IDM as in (4) with the corresponding expected counts. This can provide the necessary level of cautiousness, which is clearly advisable when coping with few data. We call this approach *imprecise EM*.[2]

On the basis of the above presented idea, we can easily achieve an imprecise quantification for an HMM, or, in other words, we can learn an imprecise HMM from any observable sequence. Note that the imprecise quantification regards only the probabilities for the hidden variables, while for the observable variables we keep the precise quantification returned by the standard EM.[3] In order to see how this technique in practice, let us consider the following simple example.

Example 1: Consider a HMM defined as in Sect. 2.1, with $N=2$, $P(X_1=1)=.5$, $P(X_{t+1}=1|X_t=0)=.7$, $P(X_{t+1}=1|X_t=1)=.5$, and a single feature $\sigma=.1$ and $\mu(X=0)=-1$, $\mu(X=1)=1$. This HMM is used to generate an observable sequence $\{o_t\}_{t=1}^{T}$. Both the standard EM and the imprecise version proposed here are used to learn the parameter $P(X_{t+1}=1|X_t=0)$. The results for different values of T are in Figure 1.

According to this example, the imprecise EM seems to manifest the behaviour we expect: for large amount of data it basically collapses into the standard EM estimates, while with fewer data a more robust estimate is provided, this corresponding to a probability intervals which generally includes the true value of the parameter (and, of course, the precise estimate). Considering the structure of the IDM as in (4) and the fact that for longer chains the expectations for the counts should increase, this should happen also in the general case.

4. Supervised Action Recognition by Imprecise HMMs

Given a collection of video sequences depicting human actions, consider a feature extraction algorithm which is "local" in time, this meaning that the features are extracted independently from each frame. On the basis of these observational data, the EM algorithm can be used to learn a HMM

[1] The missing data make the likelihood function not concave and not unimodal. Thus, generally speaking, there could be multiple maximum-likelihood estimators, and even if the EM would converge to the global maximum, the estimated parameters could be completely different from those generating the data. It seems reasonable to conjecture that this issue is particularly important when only few (incomplete) data are available.

[2] Clearly, this is just a possible, simple, approach to the generalisation towards the imprecise-probabilistic framework of the EM algorithm. A more sophisticated approach would require a Bayesian formulation algorithm, and then its imprecise extension by considering a set of priors as in the IDM.

[3] Learning imprecise-probabilistic models for continuous variables from data is a problem less studied in the literature. Despite some recent advances in this field, for the moment we prefer to put imprecision only on the hidden layer of the HMM.

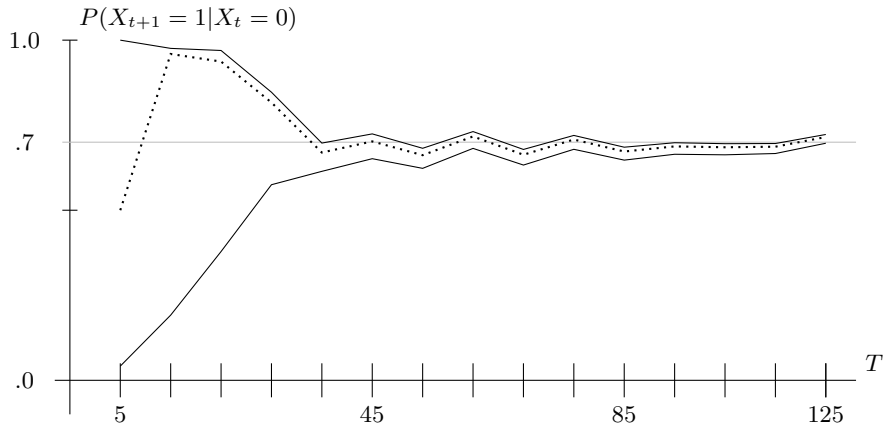

Fig. 1: Imprecise vs. precise EM. The lower and upper bounds returned by the imprecise EM (continuous lines) are compared with the precise estimate provided by the standard EM (dotted line) and with the true numerical value of the model used to generate the data (grey line).

for each sequence, and, analogously, the imprecise version of the EM proposed in Sect. 3, can be used to learn an imprecise HMM. Let $\mathcal{C}$ denote a set of (mutually exclusive and exhaustive) action labels, which can be attached to a video sequence. Given a labeled collection of sequences $\{(c_d, s_d)\}_{d=1}^{D}$, let us denote respectively by λ_d and $\overline{\underline{\lambda}}_d$ the precise and imprecise HMMs obtained from sequence s_d, and $c_d \in \mathcal{C}$ the relative action label.

To recognise a new unlabeled action s_{D+1} in the precise framework, we can simply identify the action label c_{d^*} of the HMM assigning highest likelihood to the sequence under consideration, *i.e.*,

$$d^* := \arg\max_{d=1,\ldots,D} P(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1}|\lambda_d). \quad (5)$$

To extend (5) to imprecise probabilities, we first note that, for the likelihood assigned to an observable sequence by the imprecise HMM, we can only evaluate lower and upper bounds. This corresponds to an inference (marginalisation) problem on a credal net with tree topology. The problem can be efficiently solved by the algorithm in [10], this making possible to obtain $\underline{P}(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1}|\overline{\underline{\lambda}}_d)$, and similarly the upper bound.

As the argmax in (5) only copes with point estimates, these interval data should be processed by some other optimality criterion. We choose *interval-dominance*, *i.e.*, we reject an interval if its upper bound is lower than the lower bound of some other interval. This dominance should be checked for each pair of intervals; so we end up with set $\mathcal{C}^* \subseteq \mathcal{C}$ of the classes associated to HMMs whose intervals are undominated, *i.e.*, $\mathcal{C}^*$ is the set:

$$\left\{ c_{d^*} \in \mathcal{C} \,\middle|\, \nexists d = 1, \ldots, D : \overline{\underline{\lambda}}_{d^*} \prec \overline{\underline{\lambda}}_d \right\}, \quad (6)$$

with $\overline{\underline{\lambda}}_{d^*} \prec \overline{\underline{\lambda}}_d$ meaning that:

$$\overline{P}(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1}|\overline{\underline{\lambda}}_{d^*}) \leq \underline{P}(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1}|\overline{\underline{\lambda}}_d). \quad (7)$$

The above action recognition algorithm can be regarded as a *credal* classifier, which may eventually returns more than a single candidate class for the instance under consideration. An empirical validation of this approach is in the next section.

5. Experiments and Evaluation

Our recognition algorithm is independent of the specific features extracted from the video sequence. These should be extracted at the frame level and represent information local in time (as opposed to global information as in [1]). We adopt those proposed in [12], which describes the distribution of optical flows in the whole frame as an histogram with 16 bins representing directions. Flows are computed by block-matching in adjacent frames: this approximates instant velocities, and makes such information suitable for our approach.

Simulation results on Weizmann [1] and KTH [13] benchmarks are reported in Tab. 1. With credal classifiers, the accuracy is not a sufficient descriptor of the performances. We also consider: the percentage of instances classified with a single class (*determinacy*); the average number of classes returned when the classification is indeterminate (*indeterminate output size*); the accuracy over the instances classified with a single class (*single accuracy*); the accuracy over the instances classified with more classes (*set-accuracy*). As we also have data about the precise classifier, we evaluate the accuracy of the precise method, when the imprecise is indeterminate (*credal-indeterminate precise accuracy*). The main comment to these results concerns this latter descriptor, whose low values show how the number of classes in output are an expressive indicator of the difficulty associated to the recognition of a particular action as depicted in a sequence.

Table 1: Empirical tests for the action recognition algorithm proposed in Sect. 4. We performed leave-one-out cross-validation with $N = 3$ for HMMs, and $s = 2$ for the imprecise EM.

	Weizmann		KTH	
Determinacy	84.72%	(61/72)	86.67%	(130/150)
Average Output Size	2.09	(23/11)	2.20	(44/20)
Single accuracy	70.49%	(43/61)	51.54%	(67/130)
Set-accuracy	54.55%	(6/11)	60.00%	(12/20)
Credal-Indeterminate Precise Accuracy	00.00%	(0/11)	20.00%	(4/20)

In a recent work (still under preparation), we adopt the dissimilarity measure for HMMs proposed in [14], which appears to be more informative than the likelihood, this leading higher recognition rates. The imprecise HMMs we learn by the proposed imprecise EM can be considered also in this case. With this measure, based on the HMM stationary probabilities, the comparisons between the different HMMs can be converted in points to be classified with the k-NN algorithm. In the imprecise framework, we cope with interval-valued distances, and we adopt the interval dominance criterion to perform credal classification. Yet, the discriminative analysis we performed for likelihood-based classification cannot be repeated. In fact, unlike the method based on the likelihood, in the precise case the k-NN returns a class, which is not guaranteed to belong to the set of classes returned by the imprecise approach. Table 2 depicts the results of some tests. Notably, if the classifier returns more than a single class, we are almost sure that the correct option is there. Of course, the price for such a higher robustness are higher accuracies with a significantly lower determinacy.

Table 2: Empirical validation for credal classification based on the metric in [14] extended to the imprecise model. The setup is as in Tab. 1.

	Weizmann		KTH	
Accuracy	100.00%	(72/72)	96.67%	(145/150)
Set-accuracy	100.00%	(56/56)	98.47%	(129/131)
Determinacy	22.22%	(16/72)	12.67%	(19/150)
Average Output Size	3.09	(173/56)	3.73	(488/131)

6. Conclusions and Future Work

An imprecise-probabilistic version of the EM algorithm, specialised for HMMs has been proposed and adopted for likelihood-based action recognition. The algorithm is effective in discriminating the hard-to-classify instances from the easy ones. If a single class-label is returned, this is very likely to be the correct one, while if multiple options are provided, the correct option is very likely to belong to this set. This approach seems to be particularly suited as a preprocessing tool for other action recognition algorithms. As a future work, we plan to extend to this framework other HMM-based techniques.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *In IEEE Internation Conference on Computer Vision (ICCV)*, 2005.

[2] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[3] Q. Shi, L. Wang, L. Cheng, and A. Smola, "Discriminative human action segmentation and recognition using semi-Markov model," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] L. Rabiner, "A tutorial on HMM and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[5] A. Sundaresan, A. Roy Chowdhury, and R. Chellappa, "A hidden Markov model based framework for recognition of humans from gait sequences," in *ICIP03*, 2003.

[6] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall, 1991.

[7] ——, "Inferences from multinomial data: learning about a bag of marbles," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 3–57, 1996.

[8] F. G. Cozman, "Credal networks," *Artificial Intelligence*, vol. 120, pp. 199–233, 2000.

[9] F. Lv and R. Nevatia, "Recognition and segmentation of 3D human action using HMM and multi-class adaboost," in *In European Conference on Computer Vision (ECCV)*, 2006.

[10] M. Zaffalon and E. Fagiuoli, "Tree-based creda networks for classsification," *Reliable Computing*, vol. 9, no. 6, pp. 487–509, 2003.

[11] G. M. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[12] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[13] I. Laptev and T. Lindeberg, "Local descriptors for spatiotemporal recognition," in *Proc. of ICCV*, 2003.

[14] J. Zeng, J. Duan, and C. Wu, "A new distance measure for hidden Markov models," *Expert Syst. Appl.*, vol. 37, pp. 1550–1555, 2010.