HMM-MIO: an enhanced hidden Markov model for action recognition

Oscar Perez Concha, Richard Yi Da Xu, Zia Moghaddam, Massimo Piccardi
School of Computing and Communications, University of Technology, Sydney (UTS)
PO Box 123 Broadway, NSW 2007, Australia
{Oscar.PerezConcha, YiDa.Xu, Zia.Moghaddam, Massimo.Piccardi}@uts.edu.au

Abstract

Generative models can be flexibly employed in a variety of tasks such as classification, detection and segmentation thanks to their explicit modelling of likelihood functions. However, likelihood functions are hard to model accurately in many real cases. In this paper, we present an enhanced hidden Markov model capable of dealing with the noisy, high-dimensional and sparse measurements typical of action feature sets. The modified model, named hidden Markov model with multiple, independent observations (HMM-MIO), joins: a) robustness to observation outliers, b) dimensionality reduction, and c) processing of sparse observations. In the paper, a set of experimental results over the Weizmann and KTH datasets shows that this model can be tuned to achieve classification accuracy comparable to that of discriminative classifiers. While discriminative approaches remain the natural choice for classification tasks, our results prove that likelihoods, too, can be modelled to a high level of accuracy. In the near future, we plan extension of HMM-MIO along the lines of infinite Markov models and its integration into a switching model for continuous human action recognition.

1. Introduction

In the last decade, discriminative methods have increasingly been preferred over generative approaches for the solution of classification problems. The main advantages of discriminative methods is that they are trained more closely to the objective and they do not require the training of likelihood functions which often proves difficult, especially in the case of limited samples. This trend started with object recognition and has recently become dominant also in the action recognition literature with approaches such as bag-of-features [25] [8] [15] and conditional random fields [14] [22] [29]. The features used in these methods are typically collected at each video frame as shape/motion descriptors [29] or at the so-called spatio-temporal interest points [8] [15] which are locations of discontinuity in both space and time.

Despite the dominance of discriminative approaches in the current action recognition literature, one can note persisting advantages with generative methods:

Learning class-incrementally Generative models can be learnt in an incremental fashion: wherever a new class, c_{n+1}, is to be added to a set of existing classes $\{1...c_n\}$, its learning does not require the re-learning of all the n, existing likelihoods. Instead, in the case of discriminative methods, all decision boundaries are possibly affected and required to be re-learnt.

Detection of negative-to-all classes By calling O a sequential measurement and c its corresponding class label, generative models provide explicit class-conditional likelihoods, $p(O|c)$, which can also be naturally used for tasks such as action detection by simple thresholding. This means that certain negative-to-all class observations can be detected as they receive a low likelihood in all the available classes. Conversely, when using discriminative models the class posterior $p(c|O)$ is computed only relative to the other classes. This can only explain the measurement within the known class set and cannot reject the assignment on the ground of low evidence, $p(O)$, or suggest occurrence of a previously unseen action.

Building block towards more complex graphical models More importantly, the explicit modeling of likelihood functions allows seamless integration of the generative models into larger graphical models, and hence solution of more complex problems. A useful example is that of the *switching models* [11] where multiple models are assumed to exist and only one is "selected" or "switched to" at any given time. The class-conditional likelihood at time *t*, $p(O|c_t)$, only depends on the class label and not on the time: it can then be easily combined with a prior $p(c_t|c_{t-1})$ encoding the system's dynamics of model selection. Current interest in generative models is well testified

by the recent work on hierarchical, non-parametric switching models for arbitrary number of classes from Fox, Jordan *et al*. [9].

On the other hand, despite the above theoretical advantages, it is obvious that generative models need to prove adequate accuracy over classification tasks for their likelihoods to be claimed accurate. Therefore, in this paper we present several, progressive results on the use of an enhanced HMM showing that generative models can be improved to attain classification accuracy close to that of discriminative classifiers. The enhanced HMM, named hidden Markov model with multiple, independent observations (HMM-MIO) hereafter, joins: a) robustness to observation outliers, b) dimensionality reduction, and c) processing of sparse observations. Robustness to outliers is obtained by modelling the observation densities as Student's *t* distributions [18]. Dimensionality reduction is added by using the probabilistic principal component framework [27] [24] [7] [13] [2]]. In addition, in order to deal with a variable number of observations per frame (single, multiple or none), we present simple modifications to the standard forward-backward and Baum-Welch algorithms. The modified algorithms are still in closed form and obviously efficient. Experiments are performed over two popular datasets, Weizmann [5] and KTH [25]. Although these datasets are not nearly as realistic as other more recent datasets such as Hollywood [17] or UCF50 [1], they provide the widest basis for comparison with existing work and the next sections will give evidence that they are the most suitable for the comparative aim of this paper.

The rest of the paper is organized as follows: in Section 2, we present the enhanced HMM with emphasis on its capability of processing multiple, independent observations per frame. In Section 3, we present results from two sets of experiments over the Weizmann dataset while in Section 4 we present the experimental results over KTH. Section 5 discusses future extensions and the conclusions summarise the main contributions of this work.

2. HMM with Multiple, Independent Observations

The conventional HMM is the most common generative model for time series of observations. The model expresses the joint probability, $p(O_{1:T}, Q_{1:T}|\lambda)$, of a sequence of observations, $O_{1:T} \equiv \{O_1, ...O_t, ...O_T\}$, and a sequence of corresponding hidden states, $Q_{1:T} \equiv \{Q_1, ...Q_t, ...Q_T\}$ under the well-known Markov and observation independence assumptions:

$$p(Q_t|Q_{1:t-1}, O_{1:t-1}) = p(Q_t|Q_{t-1}) \quad (1)$$

$$p(O_t|Q_{1:T}, O_{1:t-1}, O_{t+1:T}) = p(O_t|Q_t) \quad (2)$$

Each observation is typically a feature vector of fixed size. Observations belonging to a single state are commonly modelled by mixture distributions, often using the Gaussian as the basis distribution [23]. To increase robustness to outliers, also the *t* distribution has been used as basis distribution [6] [21]. In the case of feature vectors of high dimensionality, dimensionality reduction can also be easily incorporated into HMM by using the probabilistic principal component framework [27] [24] which also allows extensions to the *t* distribution. Given that typical action feature sets are affected by severe measurement noise and high dimensionality (for instance, the HOG/HOF descriptor of [17] has a combined dimensionality of 145), in this work we adopt the approach presented by [21] to simultaneously mollify outliers and dimensionality. However, the conventional HMM is not designed to deal with the variable number of observations per frame common in action feature sets. As an example, Figure 1 shows application of a STIP detector [16] to the KTH dataset and the corresponding variable number of STIP points. In order to comprehend this case, in the following we introduce a new variant of HMM, which we refer to as "HMM with multiple, independent observations" (HMM-MIO).

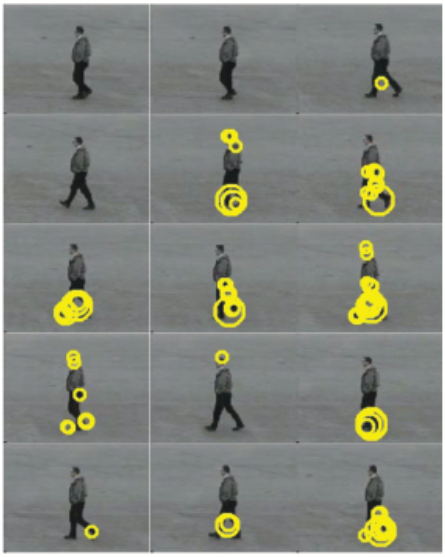

Figure 1. An example of KTH sequence, sampled at every 8 frames with STIP points plotted. Note the variable number of STIPs appearing in subsequent frames.

For the multiple observations at each video frame, we

propose the simplifying assumptions of independence and identical distribution under a mixture. By calling N_t the number of observations at time t, we define:

$$p(O_t^{1:N_t}) \equiv p(O_t^1, ..., O_t^{N_t}) = \prod_{n=1}^{N_t} P(O_t^n|Q_t), \text{if } N_t > 1$$
$$= 1, \text{if } N_t = 0 \tag{3}$$

Posing $p(O_t^{1:N_t})$ = 1 in the case of no observations is equivalent to a missing observation and has neutral effect in the chain evaluation of the HMM. We then obtain the new generative model:

$$p\left(O_{1:T}, Q_{1:T}|\lambda\right) \equiv p\left(\{O_t^{1:N_t}, Q_t\}_{t=1}^T|\lambda\right)$$
$$=p(O_1^{1:N_1}, Q_1, ..., O_t^{1:N_t}, Q_t, ..., O_T^{1:N_T}, Q_T|\lambda) \tag{4}$$

2.1. Scale of the observation probabilities in HMM-MIO

A side effect of introducing multiple observations into equation (4) is that the scale of the probability for all the observations in a frame, $p(O_t^{1:N_t})$, may vary considerably with their number, N_t. This is an undesirable effect since the number of features such as STIPs varies significantly along the frame sequence and cannot be regarded as an indicator of the reliability of the measurement process. We therefore find it desirable that the scale of probability $p(O_t^{1:N_t})$ be the same at each frame, irrespectively of the number of the observations in the frame. For this reason, we decided to normalize the probability using the following equation (5). We refer to this "scaled-consistent" probability as p^g:

$$p^g(O_t^{1:N_t}|Q_t) = \sqrt[N_t]{\prod_{n=1}^{N_t} P(O_t^n|Q_t)} \tag{5}$$

In the case of N_t = 0 (no observations in the frame), we pose $p^g(O_t^{1:N_t}|Q_t)$ = 1 so as to equate the effect of a missing observation and the absence of the corresponding edge in the HMM's graphical model. Referring to Rabiner's popular shorthand notation for the observation probability, $b_j(O_t) = p(O_t|Q_t = j)$, we extend it to $b_j^g(O_t^{1:N_t}) = p^g(O_t^{1:N_t}|Q_t = j)$ [23]. After this normalization, the HMM-MIO's generative model becomes:

$$p\left(O_{1:T}, Q_{1:T}|\lambda\right) =$$
$$=p(Q_1)\prod_{t=2}^{T} p(Q_t|Q_{t-1})\prod_{t=1}^{T} p^g(O_t^{1:N_t}|Q_t)) \tag{6}$$

In logarithmic form, as used by the expectation-maximization algorithm for the learning of parameters, the generative model with normalized scale is:

$$\ln p\left(O_{1:T}, Q_{1:T}|\lambda\right) =$$
$$= \ln p(Q_1) + \sum_{t=2}^{T} \ln p(Q_t|Q_{t-1}) + \sum_{t=1}^{T} \ln p^g(O_t^{1:N_t}|Q_t)) \tag{7}$$

with

$$\ln p^g(O_t^{1:N_t}|Q_t)) = \frac{1}{N_t}\sum_{n=1}^{N_t} \ln P(O_t^n|Q_t) \tag{8}$$

Equation (8) justifies the form taken by the update equations presented in the following Section 2.3.

2.2. Forward and backward formulas for HMM-MIO

The forward and backward formulas for the traditional HMM have been changed to accommodate the multiple observations of HMM-MIO. Following notations in [4], the forward formula, i.e., $\alpha_i(t)$, is now changed to $\alpha_i^g(t)$:

$$\alpha_i^g(t) = p^g(O_{1:t}, Q_t = i|\lambda) \tag{9}$$

The recursion in the forward algorithm is then specified as:

1. $\alpha_i^g(1) = \pi_i b_i^g(O_1^{1:N_t})$ (10a)
2. $\alpha_i^g(t) = \left[\sum_{j=1}^{R} \alpha_j^g(t-1)a_{ji}\right] b_i^g(O_t^{1:N_t})$ (10b)
3. $p(O_{1:T}|\lambda) = \sum_{i=1}^{R} \alpha_i^g(T)$ (10c)

where a_{ij} and π_i indicate the transition probabilities between any two states, and the initial probabilities, respectively. In the above equations, and in the rest of this paper, R refers to the number of possible hidden states. Like the forward formula, the backward algorithm is changed from the usual $\beta_i(t)$ to $\beta_i^g(t)$:

$$\beta_i^g(t) = p^g(O_{t+1:T}|Q_t = i, \lambda) \tag{11}$$

The corresponding recursion in the backward algorithm is then formulated as:

1. $\beta_i^g(T) = 1$ (12a)

2. $\beta_i^g(t) = \sum_{j=1}^{R} a_{ij} b_j^g(O_{t+1}^{1:N_{t+1}}) \beta_j^g(t+1)$ (12b)

3. $p(O_{1:T}|\lambda) = \sum_{i=1}^{R} \pi_i b_i^g(O_1^{1:N_1}) \beta_i^g(1)$ (12c)

Similarly, we replace the expression for the state posterior at time t, $\gamma_i(t)$, given in [4] with $\gamma_i^g(t)$, obtaining:

$$\gamma_i^g(t) = p^g(Q_t = i|O_{1:T}, \lambda) = \frac{\alpha_i^g(t)\beta_i^g(t)}{\sum_{j=1}^{R} \alpha_j^g(t)\beta_j^g(t)} \quad (13)$$

However, the posterior probability for the mixture component generating an observation must still be computed individually for each observation. Therefore, the following holds:

$$\gamma_{il}(O_t^n) = p(Q_t = i, X_{it}^n = l|O_{1:T}, \lambda) = \gamma_i^g(t) \frac{c_{il} b_{il}(O_t^n)}{b_i(O_t^n)} \quad (14)$$

where X_{it}^n is a random variable indicating the mixture component for observation O_t^n for state i, c_{il} notes the component's weight in the mixture, and $b_{il}(O_t^n)$ is the probability of observation O_t^n in the l-th mixture component for state i, $l = 1...M$.

2.3. Update equations for the observation probabilities in HMM-MIO

All observation probabilities applied in this paper are mixture models. Similarly to the work in [21], the mixture models we experimented on include the Gaussian mixture model (GMM), the mixture of principal component analyzers (MPPCA) and the mixture of t distribution subspaces. For HMM-MIO with GMM observation probabilities, the update equations are changed from the traditional HMM as follows:

$$\text{HMM:} \quad c_{il} = \frac{\sum_{t=1}^{T} \gamma_{il}(t)}{\sum_{t=1}^{T} \gamma_i(t)} \quad (15a)$$

$$\text{HMM-MIO:} \quad c_{il} = \frac{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n)}{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_i^g(t) = \sum_{t=1}^{T} \gamma_i^g(t)} \quad (15b)$$

$$\text{HMM:} \quad \mu_{il} = \frac{\sum_{t=1}^{T} \gamma_{il}(t) O_t}{\sum_{t=1}^{T} \gamma_{il}(t)} \quad (16a)$$

$$\text{HMM-MIO:} \quad \mu_{il} = \frac{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n) O_t^n}{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n)} \quad (16b)$$

$$\text{HMM: } \Sigma_{il} = \frac{\sum_{t=1}^{T} \gamma_{il}(t)(O_t - \mu_{il})(O_t - \mu_{il})^T}{\sum_{t=1}^{T} \gamma_{il}(O_t)} \quad (17a)$$

HMM-MIO: (17b)

$$\Sigma_{il} = \frac{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n)(O_t^n - \mu_{il})(O_t^n - \mu_{il})^T}{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n)}$$

2.3.1 Update equations for MPPCA and the mixture of t distributions

Probabilistic principal component analysis (PPCA) is based on the assumption that an observed sample, O, is explained by a noisy linear process over a latent variable, x, of lower dimensionality:

$$O = Wx + \mu + \epsilon \quad (18)$$

PPCA assumes that $x \sim \mathcal{N}(0, \mathbb{I})$ and $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I})$, independently [27]. Therefore, $O \sim \mathcal{N}(\mu, WW^T + \sigma^2\mathbb{I})$. This implies that O is constrained to lie in an embedded subspace if not for the effect of some, somehow small isotropic noise. In terms of parameters, Σ is replaced by W and σ^2. Maximum-likelihood solutions were derived by Tipping and Bishop in closed form [27] and Sam Roweis via expectation-maximization [24]. When considering HMM and mixtures, these parameters must be computed per state and component. The update equations become:

$$\mu_{il} = \frac{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n)(O_t^n - WE[x_{ilt}^n])}{\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n)} \quad (19)$$

where by $E[x_{ilt}^n]$ we have noted the expected value of the latent variable for state i and component l conditioned on the sample, O_t^n, and:

$$W_{il} = \left[\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n)(O_t^n - \mu_{il}) E[x_{ilt}^n]^T\right] \left[\sum_{t=1}^{T} \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{il}(O_t^n) E[x_{ilt}^n x_{ilt}^{nT}]\right]^{-1}$$

where by $E[x_{ilt}^n x_{ilt}^{nT}]$ we have noted the second-order moment of the posterior distribution of the latent variable for state i and component l conditioned on the sample, O_t^n. For brevity, we omit the update formula for σ^2 which can be easily derived from [26]. The update formulas for the mixture of t distribution subspaces follow a similar derivation, with the addition of a further latent variable for the *scale* of the generating covariance [18] [2].

3. Applying the enhanced HMM to the Weizmann dataset

This section describes two experiments where the enhanced HMM is applied to the Weizmann dataset with the use of different feature sets.

3.1. Robust HMM using dimensionality reduction

The Weizmann Institute of Science dataset consists of videos from 9 different actors performing 10 primitive actions each [5]. The action classes include walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack and skip. In this dataset, action recognition is significantly facilitated by the availability of the actors' masks in all frames. As a first feature set, we have considered an unspecialised feature set composed of all the mask's pixels. However, given that the masks vary in size over the various frames and videos, we first resized them all by re-sampling to a size of 16x16 pixels, equivalent to a feature vector with $f = 256$ dimensions. Figure 2 shows 20 frames of an action's masks from the dataset. Note that while the original masks are binary, the resized images are mildly in grey-level from the interpolation of the original binary pixels. We also note that this feature set does not entail a variable number of observations per frame.

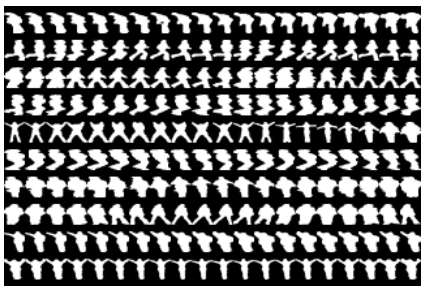

Figure 2. Example of the Weizmann action masks for one of the actors ('daria'), depicting 8 of the actions. All the masks were resized to 16x16.

We experimented with the enhanced HMM by comparing GMM (full, diagonal, spherical covariances), MPPCA and the mixture of t distribution subspaces as observation probabilities. We fixed the number of states, R, to five and the number of components per mixture, M, to two, and used leave-one-actor-out cross-validation for assessing the classification accuracy. One HMM was trained per class and classification simply provided as the class scoring the maximum likelihood for the video (class priors are all uniform for benchmark datasets). MPPCA obtained the highest average accuracy over five successive random starts with 96.9 $\pm$ 1.45%, well above the best configuration of GMM (full, with 94.0 $\pm$ 0.61%). The mixture of t distribution subspaces, with an accuracy 95.8 $\pm$ 0.93%, did not improve over MPPCA, probably due to a lack of substantial outliers in the feature set.

To demonstrate that the enhanced HMM can benefit from simultaneous dimensionality reduction and robustness to outliers, we then artificially created a noisy dataset from the original Weizmann. An 8x8 *noisy square* corresponding to 25% of the image area was added to every 3rd frame at a random position. The pixels within this noisy square switch their values from 0 to 255, or from any value different from 0, to 0. This process simulates the typical segmentation errors of foreground extraction and makes the dataset more realistic. Figure 3 shows the results of adding noise to the original data set.

Figure 3. *Noisy* Weizmann action masks for one of the actors ('daria'; the original are depicted in Figure 2). An 8x8 *noisy square* is added every 3 frames at a random position.

The experimental results on the noisy version showed that the introduced artifacts do behave as outliers of the mixture distributions. In this case, the mixture of t distribution subspaces achieved the highest accuracy, 96.2 $\pm$ 0.99%, against the 95.6 $\pm$ 0.79% of MPPCA. While the difference is not remarked, the inversion in trend with respect to the non-noisy case gives evidence to the benefit of robust models in mitigating the effects of outliers. It is also interesting to note that the accuracy of the mixture of t distribution subspaces is slightly higher on the noisy dataset (96.2 $\pm$ 0.99%) than it was on the original one (95.8 $\pm$ 0.93%). However, we are inclined to believe that this is a side effect of the addition of noise and may not bear statistical significance.

3.2. Robust HMM with a specialised feature set

The experiments in the previous subsection were conducted with a nondescriptive feature set, agnostic about the nature of the human subjects. In this subsection, we con-

sider experiments with a specialised feature set, the *sectorial extreme points* of [19], which extracts five notable points from the actor's silhouette in loose association with the head, hands and feet of the actor. The actor's centroid is added to the representation for a total of f = 12 dimensions in the feature vector. Given the low dimensionality of this feature set, we have decided to only experiment by comparing GMM and the mixture of t distributions without dimensionality reduction. In addition, instead of choosing arbitrary numbers of states and components, we have made them vary in range $R, M \in \{1...6\}$. Accuracy was again assessed over multiple randoms starts by leave-one-actor-out cross-validation. The best result achieved by GMM was 96.8%, while the best result achieved by the mixture of t distributions was 100%. Such an accuracy is obviously the highest possible and equals that reported by a few other papers to date, all based on discriminative models, including [28] which used a factorial conditional random field and [20] which made use of an ensemble of support vector machines. Moreover, the accuracy with the robust model was higher than that of GMM for 33 out of 36 combinations of R and M. These results further confirm the usefulness of the t distribution in outlier mitigation.

4. Applying the enhanced HMM to the KTH dataset

The KTH dataset is a more probing dataset than Weizmann in terms of both sheer number of videos (2,391 in total, from 25 actors) and acquisition conditions, inclusive of four different scenarios and mild camera movements. The action classes include walking, jogging, running, boxing, hand waving and hand clapping. Although KTH is becoming saturated in recent years with results reporting high accuracies, it still offers the widest platform for comparison with previous work [10]. In this paper, in order to establish a fair comparison and limit its scope to inferential methods rather than feature design, we have chosen to adopt the same features, STIPs, of a deservedly much-cited paper from Laptev *et al*. [16]. STIPs have gained increasing popularity for action recognition since they describe salient points in space and time and do not require a preliminary step of foreground extraction which is generally regarded as inaccurate. For this paper, we have used a combination of HOG and HOF descriptors [16] [17] for an overall dimensionality of 145 dimensions. The main difference with [16] is that we do not convert sets of such descriptors into sparse histograms; rather, we use each descriptor individually as an observation for HMM-MIO.

The experiments conducted were, again, performed by comparing GMM, MPPCA and the mixture of t distribution subspaces as observation probabilities. For both MPPCA and the mixture of t distribution subspaces, we evaluated over a range of increasingly reduced dimensions including 36, 18, and 9. For the mixture of t distribution subspaces we manually selected different values of the *degrees of freedom* parameter. Most of the experiments were performed with R=10 and M=5, yet not all possible combinations were tested due to the length of each experiment (approximately four hours for a Matlab implementation and an Intel Core 2 2.4 GHz PC). For evaluation, we carefully followed the procedure adopted by Schuldt *et al*. in [25]: the KTH sequences were grouped into three sets, namely, training, validation, and test, comprising of specific actors from the dataset in the number of 8, 8, and 9, respectively. The HMM-MIO were trained on the training set, one per class, and the validation set was used to select the best parameters based on maximum validation accuracy. Finally, the parameters selected from the validation set were used over the test set to provide the final accuracy results [25].

		Valid. accuracy (%)	Test accuracy (%)
GMM	Σ=full	87.3	79.7
	Σ=diag.	81.8	74.3
	Σ=spher.	79.7	72.9
MPPCA	*D=36*	87.6	82.0
	D=18	84.0	81.2
	D=9	84.3	76.6
Mt-ss (ν=3)	*D=36*	86.5	80.4
	D=18	87.3	80.4
	D=9	**91.2**	**85.7**
Mt-ss (ν=2)	*D=36*	87.7	80.4
	D=18	89.1	80.4
	D=9	90.2	84.9

Table 1. Accuracy (%) of HMM-MIO over the KTH dataset with different observation probabilities: GMM (full, diagonal, spherical), MPPCA and mixture of t distribution subspaces (Mt-ss).

Table 1 shows the results for the various combinations of observation probabilities and main parameters. The first comment is that accuracies are rather high in general, showing that the enhanced HMM can utilise individual STIP descriptors as its observations despite their sparsity in time. The conventional GMM performed worse than both MPPCA and the mixture of t distribution subspaces, and even more pronouncedly with constrained covariance matrices. MPPCA achieved a highest accuracy of 82.0% on the test set, while the mixture of t distribution subspaces achieved 85.7% (and an also remarkable 91.2% accuracy over the validation set with the same parameter values). These results are still lower than the best result reported by Laptev, 91.8% on the test set, in [17]. However, they are higher than results from other papers based on STIPs such as Schuldt *et al*. [25] (71.7%) and Dollár *et al*. [8] (81.2%). Other recent papers have reported accuracies of 97% or above over KTH, but results are not directly comparable as feature sets differ significantly [12].

5. Near-future work

Results obtained to date seem to show that by using a generative, HMM framework it is possible to achieve classification accuracies which are comparable to or only moderately lower than those of discriminative approaches based on spatio-temporal features. Such results give evidence to the accuracy and flexibility of the proposed likelihood models. However, in our near-future work we intend to explore further enhancements to the HMM. Two of the goals are:

Modeling a variable number of hidden states: one limitation of our current work is that of assuming the same number of HMM states for every action class. We consider such an assumption to be weak, as different actions may enjoy different number of hidden states when modelled using HMM, with each state in rough correspondence with a particular human pose. Anecdotally, we can see that more complex actions should be modeled with a larger number of hidden HMM states than for simple actions. In order to improve the model on this aspect, we will infer the optimal number of hidden states from the data. The starting point of this unit of work will be the infinite Hidden Markov Model (iHMM) framework [3].

Integrating the enhanced HMM into a switching model: we are inclined to believe that the HMM is more suitable for human action recognition than linear dynamical systems or auto-regressive models thanks to its "coarse", intrinsically non-linear notion of state. We therefore plan to extend the switching model presented in [9] to the HMM, while integrating aspects of dimensionality reduction and robustness into the observation likelihoods.

6. Conclusions

In this paper, we have presented an enhanced HMM capable of effectively dealing with the feature sets typical of action recognition. As evidenced in the paper, such feature sets can be high dimensional, affected by outliers and based on time-irregular observations. The proposed model, HMM-MIO (hidden Markov model with multiple, independent observations), significantly amends these issues and provides remarkable accuracy when used for action classification. The accuracy achieved over two popular action datasets, Weizmann and KTH, proved comparable to that of discriminative approaches, with a best accuracy of 100% on Weizmann and 85.7% on the KTH test set. However, we claim that the main reason for a persisting interest in generative models is their flexibility over a variety of action-related tasks such as detection, time segmentation and model selection. The proposed enhanced HMM offers a contribution to the accurate modelling of likelihoods towards such tasks.

References

[1] Ucf50, http://server.cs.ucf.edu/ vision/data.html.

[2] C. Archambeau, N. Delannay, and M. Verleysen. Mixtures of robust probabilistic principal component analyzers. *Neurocomputing*, 71(7-9):1274–1282, 2008.

[3] M. Beal, Z. Ghahramani, and C. Rasmussen. The infinite hidden Markov model. *Advances in Neural Information Processing Systems*, 1:577–584, 2002.

[4] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4:126, 1998.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision, ICCV 2005*, volume 2, 2005.

[6] S. P. Chatzis, D. I. Kosmopoulos, and T. A. Varvarigou. Robust sequential data modeling using an outlier tolerant hidden markov model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1657–1669, 2009.

[7] D. de Ridder and V. Franc. Robust subspace mixture models using *t*-distributions. In *BMVC 2003*, pages 319–328, 2003.

[8] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72. IEEE, 2005.

[9] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *IEEE Transactions on Signal Processing*, 59(4), 2011.

[10] Z. Gao, M. Chen, A. Hauptmann, and A. Cai. Comparing evaluation protocols on the KTH dataset. *Human Behavior Understanding*, pages 88–100, 2010.

[11] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. *Neural Comput.*, 12:831–864, April 2000.

[12] K. Guo, P. Ishwar, and J. Konrad. Action Recognition Using Sparse Representation on Covariance Manifolds of Optical Flow. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 188–195. IEEE, 2010.

[13] Z. Khan and F. Dellaert. Robust generative subspace modeling: The subspace *t* distribution. Technical report, GVU Center, College of Computing, Georgia, 2004.

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[15] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64:107–123, September 2005.

[16] I. Laptev and T. Lindeberg. Space-time interest points. In *the 9th IEEE International Conference on Computer Vision, ICCV 2003*, volume 1, pages 432–439, 2003.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[18] C. Liu and D. Rubin. ML estimation of the *t* distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5(1):19–39, 1995.

[19] Z. Moghaddam and M. Piccardi. Human action recognition with MPEG-7 descriptors and architectures. In *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, ARTEMIS '10, pages 63–68, New York, NY, USA, 2010. ACM.

[20] L. Nanni, S. Brahnam, and A. Lumini. Local ternary patterns from three orthogonal planes for human action classification. *Expert Syst. Appl.*, 38:5125–5128, May 2011.

[21] O. Perez Concha, R. Xu, and M. Piccardi. Robust Dimensionality Reduction for Human Action Recognition. In *2010 International Conference on Digital Image Computing: Techniques and Applications*, pages 349–356. IEEE, 2010.

[22] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:1848–1852, October 2007.

[23] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[24] S. Roweis. Em algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, 2004.

[26] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[27] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),*, 61(3):611–622, 1999.

[28] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.

[29] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:872–879, 2009.