Robust dimensionality reduction for human action recognition

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
http://www.author.org/~second

February 3, 2014

Abstract

Human action recognition can be approached by combining an action-discriminative feature set with a classifier. However, the dimensionality of typical feature sets joint with that of the time dimension often leads to a curse-of-dimensionality situation. Moreover, the measurement of the feature set is subject to sometime severe errors. This paper presents an approach to human action recognition based on robust dimensionality reduction. The observation probabilities of hidden Markov models (HMM) are modelled by mixtures of probabilistic principal components analyzers and mixtures of t-distribution sub-spaces, and compared with conventional Gaussian mixture models. Experimental results on two data sets show that dimensionality reduction helps improve the classification accuracy and that the heavier-tailed t-distribution can help reduce the impact of outliers generated by segmentation errors.

1 Introduction

Automated recognition of human actions has garnered increasing interest in recent years for its potential usefulness in video surveillance systems, human-computer interaction, multimedia annotation and other applications. A single, encompassing definition of "human action" is not possible since human actions entail varied levels of complexity and different semantics, from basic gestures up to articulated, composite actions. A noticeable trend in human action recognition research has been that of focusing on the recognition of *primitive actions* such as running, jumping, waving and similar basic actions which can be used as a dictionary for the modelling of more complex actions. Many data sets have been publicly released and intensely utilised for research, including Weizmann [5], KTH [16], HumanEva [17] and, more recently, MuHAVi [9].

The first step for recognition consists of extracting an action-discriminative feature set. Typical feature sets consist of either global o local features [14]. Global representations imply the localization of the actor(s) and their subsequent representation based on shapes, silhouettes, edges, splines or other. Local representations detect spatio-temporal interest points and use local patches around these points to compute local descriptors [10]. Since the number of local features in each frame and sequence is variable, histograms are used to convert the extracted local features to fixed-length feature sets (histogram of oriented spatial gradient (HOG), histogram of optical flow (HOF) etc) suitable for use with statistical classifiers [11]. A single histogram may be computed over the entire frame sequence, or the frame sequence may be partitioned with a temporal grid and a histogram computed over each temporal segment. Local representations have gained momentum in recent years for their strong recognition performance [11, 14].

Once the feature set for the sequence is available, the classification problem can be solved by direct classification (using conventional classifiers such as nearest neighbours, support vector machine or any others), or by temporal state models which assume that the joint probability of the measurements can be simplified by use of latent-

state dynamics. In a temporal state model, each state represents a phase of the action and the transitions between states are governed by the assumed dynamics; at their turn, measurements are explained by state-conditional likelihoods. Temporal state models have been criticised in various ways as either being too rigid or unrealistic. However, one can see that they capture the full temporal nature of the action better than any other approaches since the feature set is measured at every frame rather than once for the entire sequence or over a grid of arbitrary size. Nevertheless, a challenge to these classifiers is posed by the combined dimensionality of the feature set, say, P, and that of the number of frames, T, leading to a possible curse-of-dimensionality situation. For instance, for an action instance occurring over 10 seconds, with a frame rate of 25 fps and a feature set with $P = 100$ dimensions, the joint dimensionality $T \cdot P$ is equal to 25,000. As much as conditional independences may be assumed along both the feature space and time, the actual dimensionality remains huge. As an additional problem, the chained evaluation typical of temporal state models may fail in the presence of outliers.

In this work, we stress the importance of designing effective state-conditional observation likelihoods to overcome the aforementioned problems. We propose to use dimensionality reduction techniques to overcome overfitting of observation likelihoods. Robustness to outliers is added by using heavy-tailed distributions such as the Student's *t*. While previous work exists utilising dimensionality reduction in time-series classifiers [13, 7], to the best of our knowledge this is the first work attempting a comparative analysis in the field of human action recognition. In this paper, we use the well-known hidden Markov model (HMM) as the time-series classifier and we compare the classification accuracy achievable with various models for the observation likelihoods. In addition to the usual Gaussian mixture model (GMM), we have used dimensionally-reduced models, namely the mixture of probabilistic principal component analyzers (MPPCA) and the mixture of t-distribution sub-spaces (Mt-ss) [19, 6, 2]. Results show that dimensionality reduction helps improve the classification accuracy and that heavy-tailed distributions are effective against typical outliers. The rest of the paper is organised as follows: Section 2 describes time-series classification and the hidden Markov model, with sub-sections addressing dimensionality reduction and the selected techniques. Section 3 describes the data set and the experimental set-up, while Section 4 presents and discusses the results. Eventually, conclusive remarks are addressed relating to this work and its current extensions.

2 Classification of time series

Statistical properties of multivariate time series such as $O = \{o_1, ..., o_t, ..., o_T\}$ can be described by probability density function $p(O) = p(o_1, ..., o_t, ..., o_T)$. Yet, such a joint probability of all the observations (each, in turn, being a P-dimensional random variable) is regarded as intractable or impractical. Therefore, the joint probability is often factorised into smaller terms, and conditional independencies between observations are explored. However, conditional independencies between observations are hard to model in general and alternative models based on the notion of latent, or hidden, states have been preferred. Common models are based on two assumptions: 1) observations are mutually independent if conditioned on the states that have "generated" them, and 2) dynamics of the model is explained by the transitions between states alone, often under first-order Markov hypotheses. In addition, for action recognition, models where the states are discrete random variables with few state values are commonly adopted as they simplify analysis.

The hidden Markov model (HMM) is a prototypical example where states $Q = \{q_1, ..., q_t, ..., q_T\}$ are posited, each in correspondence with an observation. Each $q_t, t = 1...T$ is a discrete random variable taking values over a finite set, $S = \{s_1, ..., s_k, ..., s_N\}$. The parameters of the model consist of 1) the state transition probabilities, $A = \{a_{ij} = P(q_t = s_j | q_{t-1} = s_i)\}, i, j = 1...N$, 2) the observation probabilities, $B = \{b_i(o) = p(o|q_t = s_i)\}, i = 1...N$, which for our case of continuous observations are actually probability density functions, and 3) the initial state probabilities, $\pi_i = \{P(q_1 = s_i)\}, i = 1...N$. Given that the model is stationary, A and B are the same for any t. The three groups of parameters together, $\lambda = \{A, B, \pi\}$, define the HMM completely. The observation probability density functions (pdfs) are often modelled by Gaussian mixture models with a pre-determined number of components, M, as in:

$$b_i(o) = \sum_{l=1}^{M} \alpha_{il} \mathcal{N}(o|\mu_{il}, \Sigma_{il}), \quad i = 1..N. \tag{1}$$

where α_{il} is the mixing parameter, or prior, of the l-th component and μ_{il}, Σ_{il} are its mean and covariance.

The two problems we are interested in addressing with HMM are learning and evaluation. Learning of an HMM provides a set of parameters, λ, from a set of E training examples, $O_e, e = 1..E$. The most common algorithm used for this step is the Baum-Welch algorithm which belongs to the broad family of expectation-maximisation (EM) algorithms. Given a model, λ, evaluation provides a density value for sequence O, $p(O|\lambda)$, efficiently computed by forward-backward algorithms. Maximum-likelihood classification can therefore be provided based on $p(O|\lambda)$: given a number of trained models, $\lambda_c, c = 1..C$, one for each of the C classes of interest, the maximum-likelihood class is given by

$$c_{ML} = \underset{c}{\operatorname{argmax}}(p(O|\lambda_c)), \quad c = 1..C. \tag{2}$$

The above can be easily adjusted to maximum-a-posteriori or minimum Bayesian risk by addition of appropriate priors and weights.

2.1 Dimensionality reduction

In the case of high-dimensional spaces, density estimation is challenged by the relative scarcity and possible sparseness of the training data. The resulting models often suffer from little generalization capability over unseen data. A common solution to this problem is offered by dimensionality reduction techniques. Amongst them, principal component analysis (PCA) is a term of reference. PCA maps a y sample from a high P-dimensional space to a point $x = W^T(y - \bar{y})$ in a D-dimensional space, with D typically $<< P$. From x, an approximated reconstruction of y is obtained as $\tilde{y} = Wx + \bar{y}$, with the reconstruction error defined as $\epsilon = \tilde{y} - y$. The parameters of PCA are the P x D transformation matrix, W, and offset $\bar{y}$. Both parameters are learned based on a given set of training data, $Y = \{y_i\}, i = 1..N_y$: W is given by the D "largest eigenvectors" (the eigenvectors corresponding to the largest eigenvalues) of their sample covariance and $\bar{y}$ by their sample mean. This choice for W has the effect of minimising the total squared reconstruction error over the training set. Correspondingly, it maximises the sample covariance in x-space, hoping to retain useful information. However, PCA models cannot be learned with maximum likelihood or other fuller Bayesian methods due to their incomplete probabilistic formulation.

Probabilistic PCA (PPCA) amends the limitations of PCA by proposing a full probabilistic model that can be trained with maximum likelihood. PPCA assumes the existence of a latent, low-dimensional space where a point, x, is in correspondence with a y sample in the original space. The relationship between samples and latent points is given by:

$$y = Wx + \mu + \epsilon \tag{3}$$

where W is a P x D matrix describing a linear transformation and ϵ is an additive noise component. Both x and ϵ are treated as random variables and assumed normally distributed, with $p(x) = \mathcal{N}(x|0, \mathbf{I})$ and $p(\epsilon) = \mathcal{N}(\epsilon|0, \sigma^2\mathbf{I})$, and independent. It follows immediately that $p(y) = \mathcal{N}(y|\mu, C)$, with $C = WW^T + \sigma^2\mathbf{I}$ [20].

As Gaussian models are highly sensitive to outliers during training, longer-tailed distributions such as the Student's t-distribution have been used for robust modelling [12]. To join robustness with dimensionality reduction, a sub-space version of the t-distribution was proposed in [8]. The sub-space t-distribution uses the same model of PPCA with the addition of a further random variable called a *scaling* u. Probabilities for x and ϵ are given as conditional densities on u, $p(x|u) = \mathcal{N}(x|0, I/u)$ and $p(\epsilon|u) = \mathcal{N}(\epsilon|0, \sigma^2\mathbf{I}/u)$, and $p(u)$ is assumed equal to $Gamma(\nu/2, \nu/2)$ where ν is the number of degrees of freedom of the t-distribution.

It is relatively straightforward to combine multiple dimensionally-reduced models into a mixture model. The rationale for this is to obtain locally-linear models which can approximate a nonlinear manifold. When mixtures of M component distributions are considered, the single-component pdf easily extends to M individual components with mixing parameters, $\alpha_l, l = 1..M$ [4]:

$$p(y) = \sum_{l=1}^{M} \alpha_l p_l(y) \tag{4}$$

Given that closed-form solutions for the direct maximization of the likelihood are either impossible or simply

less practical, EM algorithms are commonly used for parameter estimation of mixtures [4].

Several other models for dimensionality reduction over manifolds have also been proposed such as local linear embedding and ISOMAP [15, 18]. Compared to mixture models, they have the advantage of not needing a predetermined number of components and being more flexible in the modelling of the manifold. However, they do not suit this work since they to do not define proper densities, do not obviously extend outside their training set [3] and therefore cannot be easily plugged-in in temporal state models. In the rest of this paper we focus on the mixture of PPCA and the mixture of t-distribution subspaces (Mt-ss) as observation probabilities of HMM, and compare their performance with that of the usual GMM-based HMM. The next two sub-sections sketch the basics of these two dimensionality reduction techniques.

2.2 Mixture of probabilistic principal components analyzers

Mixture of principal component analyzers are Gaussian mixtures (1) whose covariance matrix is restricted to describe a D-dimensional sub-space. Like for a general Gaussian mixture, the E step of EM requires computing the component posteriors, or *responsibilities*, for each iteration k:

$$p(l|y_i,\mu_l^{(k)},C_l^{(k)}) = \frac{\alpha_l^{(k)}\mathcal{N}(y_i|\mu_l^{(k)},C_l^{(k)})}{\sum_{h=1}^{M}\alpha_h^{(k)}\mathcal{N}(y_i|\mu_h^{(k)},C_h^{(k)})} \quad (5)$$

The M step of EM needs to maximise the expectation of the complete data log-likelihood over the components' parameters. The mixing parameters and means are provided by the following re-estimation formulas:

$$\alpha_l^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N} p(l|y_i,\mu_l^{(k)},C_l^{(k)}) \quad (6)$$

$$\mu_l^{(k+1)} = \frac{\sum_{i=1}^{N} y_i p(l|y_i,\mu_l^{(k)},C_l^{(k)})}{\sum_{i=1}^{N} p(l|y_i,\mu_l^{(k)},C_l^{(k)})} \quad (7)$$

The formulas for the update of α_l and μ_l are the same as those of a standard Gaussian mixture model (more correctly, (7) should include a term proportional to the expected value of the latent low-dimensional variable; however, such a term would tend to nullify along the iterations). Tipping and Bishop in [19] showed that matrix W_l and the noise variance σ_l^2 can be determined from the responsibility-weighted covariance matrix, S_l, (8) by standard eigen-decomposition in the same fashion as for single PPCA (10):

$$S_l^{(k+1)} = \frac{\Sigma_{i=1}^{N}(y_i-\mu_l^{(k+1)})(y_i-\mu_l^{(k+1)})^T p(l|y_i,\mu_l^{(k)},C_l^{(k)})}{\sum_{i=1}^{N} p(l|y_i,\mu_l^{(k)},C_l^{(k)})} \quad (8)$$

$$\sigma_l^{2(k+1)} = \frac{1}{P-D}\sum_{h=D+1}^{P}\lambda_{hl} \quad (9)$$

$$W_l^{(k+1)} = U_l^{(k+1)}(L_l^{(k+1)} - \sigma_l^{2(k+1)}\mathbf{I})^{1/2}R \quad (10)$$

$$C_l^{(k+1)} = W_l^{(k+1)}W_l^{(k+1)T} + \sigma^{2(k+1)}\mathbf{I} \quad (11)$$

where $L_l^{(k+1)}$ is a DxD diagonal matrix with the D largest eigenvalues of $S_l^{(k+1)}$, $U_l^{(k+1)}$ is a PxD matrix whose columns are given by the D corresponding eigenvectors, and λ_{hl} note the discarded eigenvalues. R is an arbitrary DxD rotation matrix since the model can only be identified up to an arbitrary rotation (an irrelevant detail for its use as a probability density function).

2.3 Mixture of t-distribution sub-spaces

The principal drawback of MPPCA is its sensitivity to outliers during parameter estimation, particularly for covariances. In order to mollify this problem, the mixture of t-distribution sub-spaces (Mt-ss), also known as mixture of robust probabilistic principal component analyzers, was introduced [6, 2]. Its main advantage is that the t-distribution is heavier-tailed than the Gaussian and therefore more robust. Parameter ν in the t-distribution controls the "thickness" of the tails permitting coping with outliers without translating μ or expanding Σ. The t-distribution pdf is given by:

$$St(y|\mu,\Sigma,\nu) = \frac{\Gamma(\frac{\nu+P}{2})|\Sigma|^{-1/2}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{P/2}}(1+\frac{\Delta^2}{\nu})^{-\frac{(\nu+P)}{2}} \quad (12)$$

where $\Gamma()$ is the Gamma function, $\nu > 0$ are the "degrees of freedom", P is the dimensionality of y and $\Delta^2 = (y-\mu)^T\Sigma^{-1}(y-\mu)$.

Liu and Rubin in [12] demonstrated that the maximum-likelihood parameters of a t-distribution can be obtained with EM based on the following equality:

$$St(y|\mu,\Sigma,\nu) = \int_0^\infty N(y|\mu,\frac{\Sigma}{u})G(u|\frac{\nu}{2},\frac{\nu}{2})du \quad (13)$$

where u, called the *scaling*, is a latent variable permitting reformulation of the t distribution as an infinite mixture of Gaussians over which a Gamma prior is imposed. The Gamma prior over u depends only on ν so that:

$$G(u|\frac{\nu}{2},\frac{\nu}{2}) \propto u^{\frac{\nu}{2}-1}e^{\frac{\nu}{2}u} \quad (14)$$

The t-distribution can be extended to accommodate for dimensionality reduction in a similar way to probabilistic PCA. First, $p(x)$ is defined as the prior of the D-dimensional latent variable, x:

$$p(x) = St(x|0,\mathbf{I},\nu) \quad (15)$$

Second, conditional distribution $p(y|x)$ is defined as:

$$p(y|x) = St(y|Wx+\mu,\sigma^2\mathbf{I},\nu) \quad (16)$$

where σ^2 is the variance not captured by the low-dimensional vectors and the mean of y depends on x through the PxD-dimensional matrix W.

Finally, multiple t-distribution sub-spaces can be combined in a mixture model by a tailored EM algorithm. We implemented the equations for iterative computation of the responsibilities and maximisation in α_l, μ_l, ν_l, W_l and σ_l^2 ($l = 1..M$) derived by Archambeau *et al* in [2].

3 Experiments

For the accuracy evaluation and comparison of the three proposed action classifiers (HMM with Gaussian mixture model, MPPCA and mixture of t-distribution sub-spaces as observation probabilities), we have conducted experiments over two data sets, Weizmann [5], and a version of Weizmann corrupted by segmentation errors. The data sets are briefly described in this section alongside the main results, discussing the advantages and drawbacks of each classifier.

3.1 Data Sets

The Weizmann institute dataset [5] consists of 9 different actors performing 10 primitive actions each. The 90 sequences have a size of 180x144, de-interlaced from 50 fps. The actions carried out are 'Bend', 'Run', 'Walk', 'Skip', 'Jumping Jack', 'Jump Forward On Two Legs', 'Jump In Place On Two Legs', 'Gallop Sideways', 'Wave With Two Hands' and 'Wave With One Hand'. The data set provides the original 2D human shapes or masks. Figure 1 shows 20 frames of an action's shape masks from the data set.

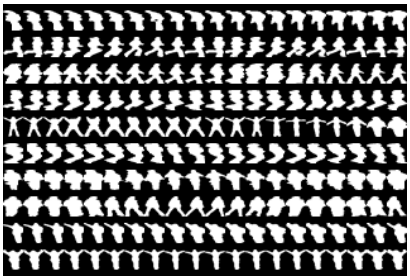

Figure 1: Weizmann actions' masks examples for one of the actors ('daria'). From first row: 'Bend', 'Run', 'Walk', 'Skip', 'Jumping Jack', 'Jump Forward On Two Legs', 'Jump In Place On Two Legs', 'Gallop Sideways', 'Wave With Two Hands' and 'Wave With One Hand'. All the masks were resized to 16x16.

The masks are resized to 16x16 pixels by re-sampling (Figure 1 shows the resized 16x16 images). While the masks are binary, the resized images are mildly in grey-level from the interpolation of binary pixels. In the next step, we construct a single 256x1 feature vector per frame by concatenating the columns of each 16x16 image. Despite its simplicity, this feature set enjoys the properties of being a) of a fixed size for all frames, b) partially invariant to antropometry, and c) independent of the subject's loca-

tion in the scene. As we conducted both training and testing from a fixed view, view invariance was not an issue in this work. However, this feature set is highly dimensional as it consists of P=256 features and is expected to create dimensionality issues. Subsequently, for each frame sequence we join its 256x1-dimensional vectors in time-frame order into an array of 256xT elements, being T the length of the sequence. The result is the input data for our HMMs. An example is shown in Figure 2 from a sequence of 20 frames (T=20) of action 'Jumping Jack'. To follow the evolution of a specific pixel during an action instance, one can select the corresponding row in the 256xT array.

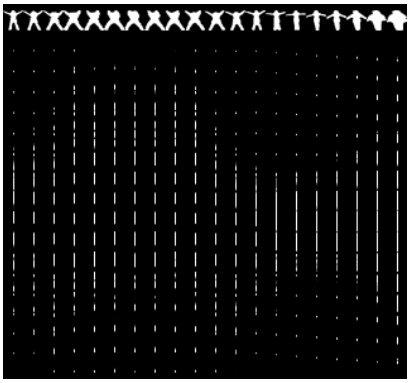

Figure 2: The picture above shows a sequence of length 20 frames (T=20) for the 'Jumping Jack' action. For every frame, a 256x1-column vector is built by concatenating each of the columns of the images. The picture shows the corresponding 256x1-column vector for each mask image.

Nonetheless, Weizmann presents ideal and clean data, leaving no space for outliers. In a real application, one expects segmentation errors to repeatedly occur due to the varying appearance of the actor and the variable background and illumination. Therefore, in order to obtain a more suitable benchmark for testing the performance of *t*-distribution over other models in the presence of outliers, we artificially created a noisy dataset from the original Weizmann. An 8x8 *noisy square* corresponding to 25% of the image area is intertwined every 3 frames 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 well the typical segmentation errors of foreground extraction and makes the dataset much more realistic. Figure 3 shows the results of adding noise to the original data set.

Figure 3: Noisy Weizmann actions' masks examples for one of the actors ('daria'). An 8x8 *noisy square* is intertwined every 3 frames at a random position. From first row: 'Bend', 'Run', 'Walk', 'Skip', 'Jumping Jack', 'Jump Forward On Two Legs', 'Jump In Place On Two Legs', 'Gallop Sideways', 'Wave With Two Hands' and 'Wave With One Hand'. All the masks were resized to 16x16.

3.2 Experimental set-up

The experiments consist of training HMMs with the three different probability density functions as observations probabilities: GMM, MPPCA and mixture of t-distribution sub-spaces (Mt-ss). In addition, the HMM-GMM is trained for the cases of full, diagonal and spherical covariance matrices. We used HMMs with 5 hidden states, 2 components per mixture, 15 training iterations, and ν=5 and 3 (reported in this order since the heaviness of the tails increases) for the case of the t-distribution. For both MPPCA and Mt-ss we tested with reduced dimensions of $D = 200$ and $D = 150$. Lower values for D were tested, but not reported in the following as they

generally led to worse results. The remaining parameters are initialized as follows:

- The initial, π, and transition, A, probabilities are initialized uniformly random.
- The mean, μ, of each component distribution is chosen randomly from within the set of training samples.
- The α weights are randomly initialized.
- All covariances for the GMM are initialized with the identity matrix, $\mathbf{I}$. In the case of the MPPCA and Mt-ss, covariances $C = WW^T + \sigma^2 I$ for all components are initialized as $W = U(L - \sigma^2 I)^{1/2} I$ where U is a PxD matrix whose columns are given by the D eigenvectors of the training data covariance, L is a diagonal matrix DxD with the corresponding D eigenvalues and $\sigma_l^2 = \frac{1}{P-D} \sum_{h=D+1}^{P} \lambda_{hl}$, with λ_{hl} corresponding to the discarded eigenvalues.

We carried out a *leave-one-actor-out* cross-validation so that the same actor will not be used for training and validation. Every actor in turn is used for validation. In addition, we repeat the whole cross-validation 5 times from different random starts in order to partially marginalise the randomness of the HMM parameters' initialisation.

Weizmann data set [5]			
		ACCURACY (%)	STD
GMM	Σ=full	94.0	$\pm$0.61
	Σ=diag.	94.2	$\pm$1.45
	Σ=spher.	93.3	$\pm$0.79
MPPCA	*D=200*	**96.9**	$\pm$1.45
	D=150	**96.0**	$\pm$1.27
Mt-ss(ν=5)	*D=200*	95.6	$\pm$1.11
	D=150	95.6	$\pm$0.79
Mt-ss(ν=3)	*D=200*	94.7	$\pm$1.45
	D=150	95.8	$\pm$0.93

Table 1: Average accuracy (%) and standard deviation for five rounds on the Weizmann data set with HMM observation probabilities: GMM (full, diagonal, spherical), MPPCA and t-distribution sub-spaces (Mt-ss). The reduced dimensions are $D = 200$ and $D = 150$.

Noisy Weizmann data set [5]			
		ACCURACY (%)	STD
GMM	Σ=full	94.7	$\pm$1.45
	Σ=diag.	93.1	$\pm$0.93
	Σ=spher.	91.3	$\pm$0.93
MPPCA	*D=200*	95.6	$\pm$0.79
	D=150	95.6	$\pm$1.11
Mt-ss(ν=5)	*D=200*	**96.2**	$\pm$0.99
	D=150	95.6	$\pm$1.11
Mt-ss(ν=3)	*D=200*	96.0	$\pm$0.61
	D=150	**95.8**	$\pm$0.93

Table 2: Average accuracy (%) and standard deviation for five rounds on the *noisy* Weizmann data set with HMM observation probabilities: GMM (full, diagonal, spherical), MPPCA and t-distribution sub-spaces (Mt-ss). The reduced dimensions are $D = 200$ and $D = 150$.

4 Discussion

Results for the original Weizmann and the *noisy* Weizmann are reported in Table 1 and Table 2, respectively. With the original dataset (Table 1), HMM with MPPCA proved the best classifier in all cases, with 96.9% (D=200) and 96.0% (D=150) average accuracy. Yet, the differences in accuracy between a full-parameter models such as full GMM (94.0%) and MPPCA were not so remarked. This was surprising to a degree as we were expecting a full GMM to experience greater difficulties in training effectively over a 256-dimensional space, especially in the scarcity of training samples. However, results with diagonal and spherical GMMs were better or equivalent, proving that the degrees of freedom of a full Gaussian model were redundant. The t-distribution sub-spaces proved to obtain higher performance than any GMM model, yet did not outperform MPPCA in any case (with performances in the interval of 94.7% and 95.8%). Since longer tails tend to "diffuse" class boundaries, this may be the cause for the increased misclassifications compared to MPPCA. We also care to add that the standard Weizmann dataset is a somehow "easy" dataset over which other authors have reported 100% accuracy in the past [1]. However, results in Table 1 are important to prove the point of this work.

When we analyse the results for the *noisy* Weizmann

dataset presented in Table 2, HMM with t-distribution sub-spaces instead obtained the best performance, with a maximum of 96.2% (D=200 and ν=5), *comparable to the best results on the clean dataset*. Therefore, t-distribution sub-spaces proved to provide a more suitable probability density model in the presence of outliers. Conversely, MPPCA and the restricted GMMs achieved remarkably worse results. In particular, the change in trend between Mt-ss and MPPCA gives evidence to the robustness of the former against the outliers.

From the experimental results, we can conclude that HMM observation probabilities based on low-dimensional manifolds can help increase accuracy of human action recognition and that longer-tailed distribution can increase robustness if the dataset is likely to contain outliers. Nonetheless, the number of reduced dimensions must be carefully chosen to secure the desired results. In addition, the t-distribution can amend the Achilles' heel of sequential classifiers i.e. the risk that the entire chained evaluation collapse to zero in the presence of even only one significant outlier. This is particularly true in high dimensions where normalised densities take on very low values and tend to underflow.

A final consideration goes to the feature set used in this work: this simple feature set was chosen as it lends itself to immediate pictorial description and intuitive analysis. Local representations such as spatio-temporal interest points [10] enjoy a number of advantages over silhouettes and pixel masks. However, in multiple-actor scenarios, the problem of associating sets of extracted spatio-temporal interest points with specific actors along the frame sequence (data association/correspondence) is obvious also for this type of descriptors. It is reasonable to expect that outliers be present regardless of how the feature set is chosen.

5 Conclusions and future work

In this paper, we have proposed performing human action recognition by HMM with dimensionally-reduced observation probabilities. Experiments have been conducted on two datasets (Weizmann and *noisy* Weizmann) comparing various probability models, including the conventional Gaussian mixture model (GMM), the mixture of probabilistic principal component analyzers (MPPCA) and the mixture of t-distribution sub-spaces (Mt-ss). The experimental results showed that, in the presence of outliers, t-distribution sub-spaces achieved the highest accuracy (96.2 $\pm$ 0.99% vs 95.6 $\pm$ 0.79% of the runner-up method, MPPCA) while in the absence of significant outliers MPPCA proved to obtain the best performance ((96.9 $\pm$ 1.45% vs 95.8 $\pm$ 0.93% of the runner-up method, t-distribution sub-spaces). These results prove that dimensionality reduction can be effective at increasing recognition accuracy and that the t-distribution is a more suitable density when the dataset contains segmentation outliers which is always likely the case in real applications. Given that the Weizmann dataset is small in size, in the immediate future we plan to extend these results to KTH and MuHAVi [16, 9], and experiment with other feature sets including histograms of special interest points.

References

[1] *Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model*, 2007. 6

[2] C. Archambeau, N. Delannay, and M. Verleysen. Mixtures of robust probabilistic principal component analyzers. *Neurocomputing*, 71(7-9):1274–1282, 2008. 2, 4

[3] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *In Advances in Neural Information Processing Systems*, pages 177–184. MIT Press, 2004. 3

[4] C. M. Bishop, editor. *Pattern Recognition and Machine Learning*. Springer, 2006. 3

[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. 1, 4, 6

[6] D. de Ridder and V. Franc. Robust subspace mixture models using *t*-distributions. In *BMVC 2003*, pages 319–328, 2003. 2, 4

[7] A. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 2, pages 681–688, 2004. 2

[8] Z. Khan and F. Dellaert. Robust generative subspace modeling: The subspace t distribution. Technical report, GVU Center, College of Computing, Georgia, 2004. 3

[9] D. I. R. C. Kingston-University. The muhavi-mas database, http://dipersec.king.ac.uk/muhavi-mas/, 2008. 1, 7

[10] 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. 1, 6

[11] 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. 1

[12] 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. 3, 4

[13] R. Pless. Image spaces and video trajectories: Using isomap to explore video sequences. In *Computer Vision, IEEE International Conference on*, volume 2, page 1433, 2003. 2

[14] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 2009. 1

[15] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000. 3

[16] 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. 1, 7

[17] L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 2006. 1

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000. 3

[19] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999. 2, 3

[20] 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. 3