Classifying Salsa dance steps from skeletal poses

Sotiris Karavarsamis, Dimitrios Ververidis, Giannis Chantas, Spiros Nikolopoulos,Yiannis Kompatsiaris
Information Technologies Institute
Center for Research and Technology Hellas
{s.karavarsamis,ververid,gchantas,nikolopo,ikom}@iti.gr

***Abstract*—In this paper, we explore building classifiers to detect Salsa dance step primitives in choreographies available in the Huawei 3DLife data set. These can collectively be an important component of dance tuition systems that support e-learning. A dance step is reasoned as the shortest possible extract of bodily motion that can uniquely identify a particularly repeatable movement through time. The representation of dance steps adopted is a concatenation of vectorized matrices involving the 3D coordinates of tracked body joints. Under this modeling context, a Salsa dance performance is seen as an ordered sequence of Salsa dance steps, requiring a multiple of the variables allocated in the representation of a single step. Following a previous work by Masurelle & Essid that discusses the classification of six Salsa dance steps from 3DLife, we show that it is possible to obtain better classifiers under a similar experimental protocol in terms of both test accuracy and F-measure. By carefully re-annotating the data in 3DLife, we refocus on the six-step classification problem and then extend the protocol to the case of 20 dance steps. In comparison to common classifiers of the trade operating on full-dimensions, we show that it is possible to produce more accurate models by computing a subspace of the data. At the same time it is possible to reduce problematic bias in resulting models due to the uneven distribution of samples across step data classes. We provide and discuss experimental findings to support both hypotheses for the two experimental settings.**

I. Introduction

Learning to dance with the help of a Kinect sensor has been a popular practice in the last 5 years. A sheer number of commercial applications in the entertainment sector have been available for the Xbox game console, selling several millions of devices. With the availability of the Kinect appliance at a low price along with the abundant supply of major open software libraries for handling and processing captured data (such as OpenNI[1]), a large number of ideas have been prototyped. A prominent example is the commercial Xbox Dance Central[2] title. The game helps human dancers learn how to perform dance gestures and choreographies, offering "more than 650 dance movements and over 90 dance routines". It also allows for one or several subjects to perform in front of the same Kinect sensor and be graded simultaneously. The game can classify the users' dance motion in real-time and prepare visual scoring feedback to them, while it basically allows for the real sensor-observed dance scene to be actuated to a virtual one. Another commercial example that is similar to the previous one is the *Just dance* game title developed by Ubisoft.

From an academic point of view, some research effort in dance analysis has been spent in developing algorithms that can recognize how well a user can imitate certain motion patterns presented by an avatar. Several dance flavors have recently been the topic of research, such as ballet dance [1], Tsamiko dance (see [2], [3], [4]), and Salsa dance [5] that this paper deals with.

The inherent difficulty in modeling continuous motion has ties with: a) the problem of computing a reliable degree of belief about how well a realized motion resembles an ideal motion primitive; b) capturing failed attempts of movements and backtracking immediately; and c) classifying movements which are largely corrupted (e.g., a movement is performed well for some fraction of its ideal duration but it is irregular in the rest of it); d) dealing with position and scale artifacts possibly in combination with (a)-(c). Humans are able to recognize the identity of dance steps even under the existence of time lag (slow or fast performance) or scale artifacts, for example. Computationally, such artifacts can be addressed by an algorithm that aligns an observed motion trajectory signal with the signal of an ideal motion trajectory.

Motion activity in Salsa dance is naturally reasoned in terms of steps, and steps are structured as consecutive sub-steps that are in general synchronized with music beat. One sub-step regularly lasts for one quarter of the duration of a step. Sub-steps in dance motion are ideally mapped to quarters in the music meter, although this requirement can be adhered with difficulty from non-experts. Hence, if we assume that a performance is executed in order of the correct steps, then observed trajectory structure can be concretely analyzed in terms of a predefined set of ideal steps. In practice, however, this temporal analysis is inherently affected by bodily-induced signal transformations. This difficulty may be due to a dancer missing some of the steps, or the difficulty of encoding these transformations within the model.

In this paper, we are interested in learning classifiers that can recognize steps of Salsa dance from skeletal poses, i.e., algorithms that make non-soft decisions about the identity of a step. A Salsa step is sensed as the maximum possible extract of a motion trajectory that is lengthy enough to describe a commonly perceived, repeatable motion. We represent dance steps as lists of consecutively concatenated skeleton poses cap-

[1]http://structure.io/openni
[2]http://www.harmonixmusic.com/games/dance-central/

978-1-4673-8695-1/16/$31.00

tured by the technique[3] of Shotton et al. [6] which estimates the position of skeletal joints from sequences of depth images.

To prepare Salsa dance step training data from skeletal motion trajectories pertaining to performed choreographies, we have developed an annotation tool called DanceAnno. We open-source this software in hopes that it will be useful for researchers working on the same dataset or similar ones. The tool allows a user to annotate motion trajectories by letting them place starting and ending marks to delimit a compact activity. Along with a set of body joint coordinates being plotted on a time-line, the tool provides the annotator with sample-level imagery depicting the actual subject performing the activity to be annotated. Handling more modalities within DanceAnno is possible by extending its object-oriented interface with add-on application logic.

The paper of Masurelle & Essid [5] constitutes the basis of our work, which extends its findings and supplements the problem toolkit with a careful data annotation and a data annotation tool. In [5] the authors explore the applicability of generative machine learning models, namely Gaussian mixture models and Hidden Markov Models (HMMs), in predicting six Salsa dance steps from the 3DLife data set [7].

In this paper, our contributions are the following:

a) We developed an open-source, extendable dance motion annotation tool called DanceAnno that allows a user to produce annotations of events in alignment with motion trajectory signals and sample-by-sample footage of the entire activity; b) Using DanceAnno, we carefully created an accurate annotation of the performance trajectories in 3DLife to step level by filtering out noise being uninformative to the actual payload of the motion; c) We reproduce and further extend the experimental protocol of Masurelle & Essid both for the case of six steps studied (for both male-only and male-plus-female step instances), and provide a discussion on the obstacles met in classifying 20 classes using the data produced from our annotation; and d) We provide evidence that performing feature extraction with PCA on dance steps helps to create more accurate[4] classifiers, and can control the effect of class imbalance in resulting classifiers due to the uneven distribution of training samples across all of the available classes.

The classification models discussed in this paper can be a useful component within e-learning Salsa dance tuition games which may provide, among other features, visual feedback to users about the steps they perform.

In Section II, we provide a review of recent related work on human action recognition and accompanied motion descriptors. In Section III, the data set and annotation tool used are described. Section IV provides a description of the pre-processing and the dimensionality reduction algorithms used on the raw motion trajectory signals. The classification algorithms are outlined in Section V. Experimental results are given in Section VI. Section VII states conclusions and future work.

[3]The algorithm is wrapped in OpenNI2.

[4]We compute the average test accuracy and F-measure by rotating over a unique testing dancer out of a total of 14 ones that are kept for training classifiers.

II. Related work

The dimensionality of raw skeletal motion extracts over a number of k frames is $\mathcal{O}(mk)$, where m is the number of variables to describe a single skeletal pose and k is the count of frames accumulated within a contiguous motion segment. Even for small values of k (assuming m is fixed), the total count of variables that describe a succession of poses can become large to an unnecessary extent. In previous work by Bashir et al. [8] it was shown that a parsimonious representation over only a few of the total variables can be uncovered by projecting raw motion data onto a subspace spanned by a projection matrix of low rank. We come back to this hypothesis in the experimental section of this paper.

Being a critical point in the representation of motion, skeletal data in the Human Action Recognition loop can be broadly divided in the set of approaches which attempt to directly learn feature representations off from raw data and to those approaches which explicitly map raw data into a feature space governed by a feature transformation procedure. Such a procedure is generally called a *descriptor* in the literature. In this paper, we explore the case where a classifier is learned on the raw motion data without applying a descriptor.

In the rest of this paper, we will be using the shorthand notation HAR to refer to the field of Human Action Recognition. In the course of learning more complex data structure from either raw or descriptor-piped data, the area of HAR has unquestionably drawn reincarnation from prior but recent advances in image modeling for image retrieval in multimedia databases. The *bag of visual words* model [9] is a common example that has been studied extensively.

Evangelides et al. [10], [11] propose the *skeletal quad* descriptor that embeds skeletal poses into a feature space of 6 dimensions, instead of using the raw coordinate data of the skeletal pose. In addition to compressing skeletal poses, the authors show that the descriptor is invariant to basic geometric transformations. The descriptor is applied in an action recognition scenario.

Xia et al. [12] propose the Histogram of 3D Joints (HOJ3D) descriptor. The feature description technique first transforms the three dimensional coordinate space to a polar one (non-linearly), and the polar coordinate values are used to compute a 2D radial-angular discrete histogram.

Raptis et al. [13] develop a complete real-time dance gesture recognition system which has been estimated to be able to recognize $96,9\%$ of gestures over a time frame of four seconds. In their design, the authors compute the angular difference among joint pairs that are either deemed as first-order joints or second-order ones exclusive. All angular divergences are packed into a feature vector, which is the proposed skeletal pose descriptor.

Thanh et al. [14] describe an analytical technique for skeleton based HAR. At first, they divide an entire motion timeline into temporal segments. Among the temporal segments,

key skeletal poses that optimize a local objective function are selected. In that way, one action is considered as the collection of skeletal action poses. To represent the coordinate data in the histogram, the authors make use of a histogram of 3D skeletal sequences that hashes joint points into the bins of a histogram.

Yang et al. [15] discuss a natural idea for HAR and action representation introduced as the *EigenJoint* skeletal data descriptor, which attempts to learn significant patterns of motion using PCA as the data description technique. Before identifying significant components in the space of skeletal poses, the affinity matrix of skeletal joints is considered. The descriptor attempts to take into account information that reveals temporally varying activity (e.g., acceleration and velocity). The authors test their descriptor on the so-called Naive-Bayes Nearest Neighbor (NBNN) classifier in order to categorize human action sequences into predefined classes.

III. Data preparation

Fig. 1. An instance of the DanceAnno motion activity annotation tool showing one dance tutor performing a Salsa choreography.

In this section, we introduce the DanceAnno motion trajectory annotation tool which we offer publicly[5] as open-source software. As is, the tool allows a user to generate annotation for the motion trajectory data of the 3DLife data set and the Calus dance data set of the i-treasures[6] project. However, the software can be adapted to work with other data sets as well without requiring excessive changes to the original code base. The tool makes use of the object oriented programming interface of the Python 3 programming language.

To begin a new annotation, two fields are necessary to be filled within the graphical user interface of DanceAnno: a) the path to skeleton trajectories in .mat format, or in .skel (Microsoft) v2 format; and b) the path to a directory containing timestamped video frames in PNG or JPEG format. Optionally a modal text file with music beat labels can be loaded to possibly help increase the precision of the annotation and better enhance the reasoning of the human annotator. The output of the tool is a text file with the labels and the endpoint timestamps of the labels, assuming the annotated events do not overlap temporally. The user has the option to select which joints should be plotted by the visualizer (e.g., because some signals may be less worthwhile to plot due to them being uninformative). The 'left foot' and 'right foot' trajectories are selected by default, since they are the most informative for dance motion. Next, as shown in Figure 1, each of the 3 x/y/z coordinates of each joint are plotted in a separate canvas, resulting in 6 canvases for the default joints. The time-line can be slided with a mouse drag while visual footage is also depicted.

[5]DanceAnno is available at https://github.com/MKLab-ITI/DanceAnno

[6]EU FP7 project i-treasures http://i-treasures.eu/

TABLE I
Frequency of dance steps from 14 dancers in 3DLife.

Class	Name	Occurences
1	forward-basic-step	143 (123 in [5])
2	backward-basic-step	71 (144 in [5])
3	right-turn	1
4	cross-body	1
5	preSuzieQ-backward-step	24
6	suzie-q-a	28 (18 in [5])
7	suzie-q-b	28 (18 in [5])
8	preDoubleCross-backward-step	24
9	double-cross-a	25 (18 in [5])
10	double-cross-b	26 (18 in [5])
11	pachanga-tap-a	9
12	pachanga-tap-b	9
13	swivel-tap-a	9
14	swivel-tap-b	7
15	cross-body-a	9
16	cross-body-b	9
17	turn-call	14
18	girl-turn	14
19	boy-turn-hand-switching	12
20	girl-turn-caress	12
21	cross-body-c	9
22	cross-body-d	10

A. *Sufficiency of data population*

A basic problem when building classification models for modeling aspects of Salsa dance is that often there is a limited availability of data. Growth in the amount of samples captured for the same event (in our case, a dance step or a whole choreography) occurs very slowly. Therefore, at any given point in time a model may be required to be learned using only the data available. Often it can be the case that some classes of an event are sufficiently populated while other classes are very poorly populated. In such a case, we may face the so-called *class imbalance* problem [16]. When training classifiers on class-imbalanced data, the learned model may suffer from a problematic bias in future predictions on test data. Most often a model may be observed to classify input data to only a very limited set of classes erroneously.

In this paper, we target the class imbalance problem in the classification of six and twenty steps. To lessen the class imbalance effect on the learned models in both cases we reduce the dimensions of the data with linear PCA. This transform

proves to increase the average testing accuracy, and average F-measure (with leave-one-out model testing), while also taking account of the bias in learned classifiers. We discuss this behavior in Section VI by comparing learned models on the original step feature space and a lower dimensional space determined by linear PCA.

IV. Raw signal transformations

The transformations applied to the raw signals before a model is adapted to them are linear interpolation, normalization and dimensionality reduction. Linear interpolation was used in order to make all steps of similar length (specifically, of 32 signal samples at a sampling rate of 25 Hz). Zero-mean normalization is applied to the data in order to bring them to a space of similar scale and position, since the distance between the subject and sensor, and also the subject's height, may affect the predictions cast by the models. At validation time, we can use the normalization constants that we learned at training time, as an approximation of the right position and scale of the features.

For zero-mean normalization, we compute the mean and standard deviation vector (μ and σ) of the steps in the aggregate set $\mathcal{X}$ of steps on D dimensions. For each step feature vector $\mathbf{x} \in \mathbf{R}^D$, we compute a normalized instance of it using the simple formula

$$y_i = \frac{x_i - \mu_i}{\sigma_i} \text{ for all } i \in [1, ..., D], \tag{1}$$

hence obtaining a new vector $\mathbf{y} \in \mathbb{R}^D$ in a feature space confined by the scale[7] in the data from which the normalization parameters where derived. x_i and y_i are the $i-$th scalar components of the signals $\mathbf{x}$ and $\mathbf{y}$, respectively. Therefore, a feature vector over 32 temporal moments $\times$ 15 joints $\times$ 3 x/y/z coordinates is obtained, which amounts to a total of 1440 dimensions.

In previous work, Masurelle & Essid [5] use only the skeletal joints below the waist of a subject under the assumption that over-the-waist joints are uninformative in the context of Salsa dance. In our experiments, we only observe a small increase in test accuracy when male-and-female data are used for classification. Experiments are carried out on six and twenty step classes, on male-only and male-and-female data, using six lower-body joints and fifteen whole-body joints. In a secondary data pre-processing step, the raw signals are projected onto a lower dimensional subspace computed by linear PCA (see [17]).

Limitations of the hypothesis The classification models tested on step data from 3DLife [7] are learned and validated in a controlled manner, in the sense that the length of step extracts is fixed. As step data exhibit variance generated by the motions of different subjects, the pipeline discussed does not take into account special transformations such as time lag or other spatiotemporal distortions. In this way, we assess how well the model can penalize step data that cannot express the step content within a fixed time frame. Such transformations would also place a requirement to do some form of signal registration with a dynamic time warping algorithm.

V. Classification methodology

Three classifiers have been employed in order to assign labels to steps: Random Forest models, linear and non-linear support vector machines (SVMs), and multi-class AdaBoost (called AdaBoost.M2 [18]) using decision trees as a weak learners.

A. Motivation for random forest models

The Random Forest model is a majority voting classifier that collectively considers the partial assertions of a series of decision trees. Each decision tree can consider a subset of the total variables used in a data set. After computing a model, an input datum is passed through a succession of decision trees. Each tree casts a vote for the input, and the label voted by the maximum number of decision trees is predicted.

B. Motivation for Support Vector Machines

Support Vector Machines (SVM)s is a family of classifiers [19] introducing the notion of the maximum margin separating hyperplane. It has been the core or basis of many successful applications and algorithms in machine learning. In the simple case where the data are linearly separable points on the plane, a linear SVM finds a hyperplane (in this case, a line) whose margin (the free space between the separating hyperplane and a line passing through at least two points of the positive and negative class, alike) is maximum. The parameters of the line are found by solving a quadratic optimization problem. The linear case can be extended to a non-linear setting where distance between any two points in feature space is warped non-linearly. In this paper, we consider one-versus-all (OVA) multi-class construction on n classes, where each SVM model is trained on the i-th (positive class) and the rest of $n-1$ classes (negative class). The one-versus-all case for SVM models has been studied extensively in the work of Rifkin [20].

VI. Experiments

In this section, we explore the ability to learn classifiers that can discriminate multiple Salsa dance steps. We measure the effect of class imbalance with the discrete entropy of the normalized misclassification histogram, i.e., as $-\sum_{i=1}^{C} c_i log_2(c_i)$, where c_i is the proportion of misclassified samples as belonging to class i, and C is the count of classes. A misclassification histogram is computed over all of the testing samples across 14 training rounds.

Comparison with Masurelle & Essid We cross-compare the six class classification results reported in [5]. The best F-measure[8] reported by Masurelle & Essid is attained by a Hidden Markov Model, and it is estimated at 74%. In Table I we depict the frequency distribution of steps in both the

[7] If $\sigma_n^i = 1$, we get positional normalization that does not affect scale.

[8] F-measure is defined as $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$.

authors' paper and also considering our annotation of noise-free steps in 3DLife. Using data emerging from our annotation, we find a linear SVM in a one-versus-all (OVA) scheme on 101 dimensions (amounting to 7.01% of the total variables) that attains an average F-measure of 84.05% under a variance of 12.60%, which gives the best observed performance higher by 10.05% from the best baseline score (experiment 2). The corresponding average test accuracy is estimated at 93.12% for this classifier. The previous result is attained when 6 skeletal joints are utilized. By reconsidering the six class problem this time on 15 joints, we arrive at a slightly better model on 41 dimensions with an F-measure of 84.34%. This validates the argument of Masurelle & Essid that above-the-hip skeletal joints are uninformative for the classification task. Experiment 5 shows that misclassification histogram entropy is better than the case where full dimensions are used by at least 0.21 (see Figure 2(i) and Figure 2(j)).

Our extensions We now focus on the same experimental protocol, but this time we consider the case of 20 classes[9] from choreographies c3, c4 and c5 in 3DLife. We consider our manual annotation of dance steps, aspects of which we described earlier. The best model found in this experiment is a random forest operating on a reduced subspace of 51 dimensions. The model on reduced dimensions attains an average test accuracy of 67.12% and F-measure of 44.91%. The higher the test accuracy and F-measure the more robust the resulting model. This result improves the classification accuracy of the same model in full dimensions, in which case the test accuracy is 54.79% (lower by 12.33% than the previous result) and the F-measure is 29.70%. In Figure 2(e) and Figure 2(f), the misclassification histograms on 1440 dimensions and 51 dimensions are shown. In the second histogram, we get a better entropy value estimated at 4.34 (while the value of the entropy on full dimensions is 4.19), validating our hypothesis. For the case of one-versus-all support vector machines, the best PCA projection on 34 dimensions yields a classifier with an F-measure that is slightly worse than the one computed on 1440 dimensions. The F-measure of the latter is at 23.28% while that of the former is at 21.66%. However, the misclassification histogram of the classifier on 34 dimensions has an entropy value of 3.55, which is only slightly better than the entropy of the same model on full dimensions that equals 3.52. Similarly, a marginal increase in entropy for the misclassification histogram of an AdaBoost.M2 model for the case of 3 dimensions is also observed. When PCA is used we get an entropy value of 4.14, while in the case of full dimensions the entropy value is 4.10. For reference on the misclassification histograms of classifiers operating on 20 step data classes, see Figure 2(a) through to Figure 2(d).

Entropy of misclassification histograms The class imbalance problem [16] can severely affect the applicability of classifiers. Usually, this is manifested by the classifier consistently predicting a subset of possible classes. Figure 2(g) shows an example where a classifier trained on full dimensions falsely predicts classes 3, 4 and 5 when evaluated on the testing set. We measure this artifact by the discrete entropy of the misclassification histograms on test samples. The higher the entropy of the misclassification histogram of a classifier, the smoother the distribution of errors across classes. Figure 2(a) through Figure 2(l) demonstrate that this effect can be resolved in many models.

[9]Note that we have annotated a total of 22 steps across 3DLife. Two step classes contain only one data sample each (namely, steps *right-turn* and *cross-body* which belong to choreography c2). We exclude these two step classes from our experiments.

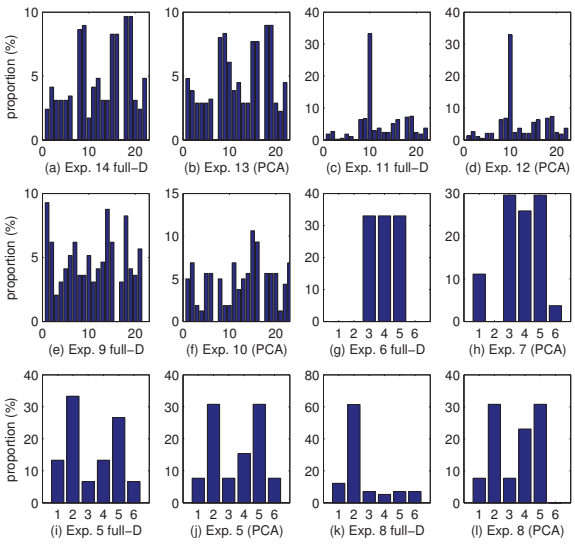

Fig. 2. Histograms show the distribution of misclassified test samples to target classes by trained classifiers with or without PCA. Odd-ordered columns show performance on full dimensions; even-ordered columns show performance with linear PCA. Tables II and III enlist the corresponding entropies computed for respective misclassification histograms (see numbering in the *Ref. #* column of each table). Histograms on full dimensions tend to accumulate a high fraction of errors to specific classes. When PCA is applied, misclassifications are distributed more evenly.

VII. Conclusions and Future Work

We conduct experiments on the classification of Salsa dance steps from the 3DLife data set. We extend the experiments of Masurelle & Essid [5] for the case of 6 step classes, and also study the case of 20 step classes. In both schemes, we assess the hypothesis that preprocessing the step data with linear PCA can help uncover a parsimonious representation that boosts the accuracy of models and tackles class imbalance. The discrete entropy of misclassification histograms shows that model validation gets better at predicting targets. In future work, we intend to integrate the classifiers explored here within in a real-time Salsa dance e-learning system which will provide real-time motion-specific feedback to users.

VIII. Acknowledgement

The research leading to these results has received funding from the European Community Seventh Framework Programme (FP7-ICT-2011-9) under grant agreement no FP7-ICT-600676 "i-Treasures: Intangible Treasures - Capturing the

TABLE II
RESULTS OF EXPERIMENTS ON 6 CLASSES IN 3DLIFE.

ref. #	problem	classifier	data	parameters	D	D/R	CCR (%)	F-measure (%)	Ent.
1	6 step [5]	Gaussian Mixture Model	M / 6J	4 components	40	PCA	n/a	68%	n/a
2	6 step [5]	Hidden Markov Model	M / 6J	3 hidden states	60	PCA	n/a	74%	n/a
3	6 step [8]	Gaussian Mixture Model	M / 6J	6 components	80	PCA	n/a	61%	n/a
4	6 step [8]	Hidden Markov Model	M / 6J	4 hidden states	40	PCA	n/a	63%	n/a
5	6 step	linear SVM / OVA	M / 6J	ℓ_2 reg/loss; $C = 0.001$	101	PCA	93.12 ± 04.87	84.05 ± 12.60	2.31
6	6 step	Random Forest	M / 15J	200 trees; mtry=$\lceil\sqrt{1440}\rceil$	1440	raw	80.35 ± 07.86	59.33 ± 11.61	1.58
7	6 step	Random Forest	M / 15J	200 trees; mtry=$\lceil\sqrt{38}\rceil$	38	PCA	91.42 ± 04.34	80.14 ± 10.83	2.31
8	6 step	linear SVM / OVA	M / 15J	ℓ_2 reg/loss; $C = 0.001$	41	PCA	93.34 ± 05.19	84.34 ± 13.20	2.10*

Abbreviations: M stands for male, J stands for joints, D stands for dimensionality; D/R stands for dimensionality reduction; CCR stands for Correct Classification Rate; Ent. stands for discrete entropy; * Entropy in the case of 1440 dimensions is 1.82

TABLE III
RESULTS OF EXPERIMENTS ON 20 SALSA DANCE STEP CLASSES FROM 3DLIFE.

ref. #	problem	classifier	data	parameters	D	D/R	CCR (%)	F-measure (%)	Ent.
9	20 step	Random Forest	M+F / 15J	200 trees; mtry=$\lceil\sqrt{1440}\rceil$	1440	raw	54.79 ± 15.58	29.70 ± 13.56	4.19
10	20 step	Random Forest	M+F / 15J	200 trees; mtry=$\lceil\sqrt{51}\rceil$	51	PCA	67.12 ± 10.52	44.91 ± 10.21	4.34
11	20 step	linear SVM / OVA	M+F / 15J	ℓ_2 reg/loss, $C = 0.001$	1440	raw	18.01 ± 07.45	21.66 ± 08.78	3.52
12	20 step	linear SVM / OVA	M+F / 15J	ℓ_2 reg/loss, $C = 0.001$	34	PCA	17.54 ± 04.30	23.28 ± 05.32	3.55
13	20 step	AdaBoost.M2	M+F / 15J	500 trees; mtry=$\lceil\sqrt{3}\rceil$	3	PCA	32.91 ± 14.50	22.03 ± 12.18	4.14
14	20 step	AdaBoost.M2	M+F / 15J	500 trees; mtry=$\lceil\sqrt{1440}\rceil$	1440	raw	36.29 ± 11.94	28.31 ± 08.01	4.10

Abbreviations: D stands for dimensionality; D/R stands for dimensionality reduction; CCR stands for Correct Classification Rate; Ent. stands for discrete entropy.

Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures".

REFERENCES

[1] M. Kyan, G. Sun, H. Li, L. Zhong, P. Muneesawang, N. Dong, B. Elder, and L. Guan, "An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment," *In ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 2, p. 23, 2015.

[2] G. Chantas, A. Kitsikidis, S. Nikolopoulos, K. Dimitropoulos, S. Douka, I. Kompatsiaris, and N. Grammalidis, "Multi-entity bayesian networks for knowledge-driven analysis of ich content," in *Proc. 1st International Workshop on Computer vision and Ontology Applied Cross-disciplinary Technologies in conj. with ECCV*, pp. 355–369, Springer, 2014.

[3] A. Kitsikidis, K. Dimitropoulos, E. Yilmaz, S. Douka, and N. Grammalidis, "Multi-sensor technology and fuzzy logic for dancerÂćs motion analysis and performance evaluation within a 3d virtual environment," in *Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access*, pp. 379–390, Springer, 2014.

[4] A. Kitsikidis, K. Dimitropoulos, S. Douka, and N. Grammalidis, "Dance analysis using multiple kinect sensors," *In Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2014.

[5] A. Masurelle, S. Essid, and G. Richard, "Multimodal classification of dance movements using body joint trajectories and step sounds," in *Proc. 14th International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1–4, IEEE, 2013.

[6] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *In ACM Journal of Communications*, vol. 56, no. 1, pp. 116–124, 2013.

[7] S. Essid, X. Lin, M. Gowing, G. Kordelas, A. Aksay, P. Kelly, T. Fillon, Q. Zhang, A. Dielmann, V. Kitanovski, *et al.*, "A multi-modal dance corpus for research into interaction between humans in virtual environments," *In journal of Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 157–170, 2013.

[8] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, "Object trajectory-based activity classification and recognition using hidden markov models," *In IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[9] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Proc. Workshop on statistical learning in computer vision in conj. with ECCV*, vol. 1, pp. 1–2, Prague, 2004.

[10] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *In Proc. ICPR*, pp. 4513–4518, IEEE, 2014.

[11] G. D. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in *In Proc. ECCV workshops*, pp. 595–607, Springer, 2014.

[12] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *In Proc. CVPR Workshops*, pp. 20–27, IEEE, 2012.

[13] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *In Proc. ACM SIGGRAPH / Eurographics symposium on computer animation*, pp. 147–156, ACM, 2011.

[14] T. T. Thanh, F. Chen, K. Kotani, and B. Le, "Extraction of discriminative patterns from skeleton sequences for accurate action recognition," *In Fundamenta Informaticae*, vol. 130, no. 2, pp. 247–261, 2014.

[15] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints," *In Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2–11, 2014.

[16] H. He and E. A. Garcia, "Learning from imbalanced data," *In IEEE TKDE*, vol. 21, no. 9, pp. 1263–1284, 2009.

[17] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," vol. 27, no. 2, pp. 83–85, 2005.

[18] Y. Freund, R. E. Schapire, *et al.*, "Experiments with a new boosting algorithm," in *In Proc. of ICML*, vol. 96, pp. 148–156, 1996.

[19] V. Vapnik, *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[20] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *In Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.