HANDS, FACE AND JOINTS FOR MULTI-MODAL HUMAN-ACTION TEMPORAL SEGMENTATION AND RECOGNITION

Bassem Seddik, Sami Gazzah and Najoua Essoukri Ben Amara

SAGE laboratory, National Engineering School of Sousse, University of Sousse, Tunisia
Email: bassem.seddik.tn@ieee.org, sami_gazzah@yahoo.fr, najoua.benamara@eniso.rnu.tn

ABSTRACT

We present in this paper a new approach for human-action extraction and recognition in a multi-modal context. Our solution contains two modules. The first one applies temporal action segmentation by combining a heuristic analysis with augmented-joint description and SVM classification. The second one aims for a frame-wise action recognition using skeletal, RGB and depth modalities coupled with a label-grouping strategy in the decision level.
Our contribution consists of (1) a selective concatenation of features extracted from the different modalities, (2) the introduction of features relative to the face region in addition to the hands, and (3) the applied multilevel frames-grouping strategy. Our experiments carried on the Chalearn gesture challenge 2014 dataset have proved the effectiveness of our approach within the literature.

Index Terms— human action recognition, temporal segmentation, Chalearn gesture challenge, Kinect, SVM.

1. INTRODUCTION

Human-action recognition is still an open research field in computer vision community. After a first trend of spatio-temporal feature extraction and learning from RGB video streams, the actual orientation leans towards the description from more than one modality combined with massive learning strategies [1]. Joints and depth streams came to enrich the existing learning inputs, and the introduction of Microsoft's Kinect sensor accelerated this progress.
The Chalearn Gesture Challenge (CGC) has been focusing in the last years on the recognition of human sign language in different scenarios such as one/multiple-shot(s) learning, single/multiple actors and variation of learning modalities [2, 3].
The present work is based on the experimental sets of the CGC 2014 dataset. It offers 4 kinds of modalities: the RGB, the depth, the mask video streams and the skeletal joints extracted from the kinect sensor. As shown in Fig.1, our goal is to automatically retrieve the motions and to recognize them from the unlabelled continuous data streams offered. The main challenges include the presence of continuous actions in addition to extra non-vocabulary actor behaviours.

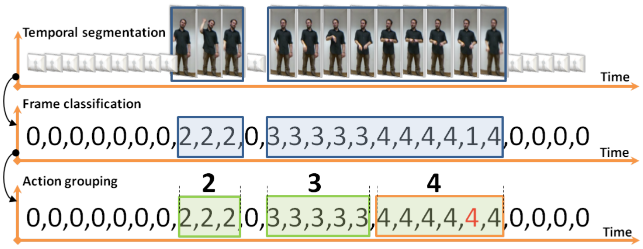


Fig. 1: Our solution first extracts action positions, then recognises each frame label. Finally, it applies a grouping strategy to get each segment's unified action-label raging from 0 to 21.

We will start in the next section by presenting the literature related to the CGC dataset, then we will give the details of our pipeline. The third section will give the details of our temporal segmentation approach. Section 4 will present our action recognition and grouping strategy. Experiments and results will be shown in section 5 and conclusion in the end.

2. RELATED WORKS

We analyse each of the existing solutions related to our context on two levels: their Temporal Segmentation (TS) solution and their approach for action recognition. Afterwards, a proposition of an improved approach is presented.

TS: A first family of solutions used a heuristic derived analysis for the TS purpose. Peng *et al.* [4] searched for the most frequented hand-joint position using a 100x100 grid. Then, they extracted any motion with a position further than a threshold T. Liang *et al.* [5] proposed a similar solution that compares between the left and right hand joint-motion and simultaneously decides about the dominant one.
Another family of solutions were derived from the dynamic programming approach. Whereas dynamic time warping has been widely applied in the past years [2], more recent works focused on the use of conditional random fields [6] and Markov random fields [7].
A last solutions family used binary classifiers to extract action regions. Evangelidis *et al.* [8] used a Support Vector Machines (SVM) classifier related to a joint energy measure. Similarly, Neverova *et al.* [9] exploited the state-of-art win-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

ning Deep Neural Network (DNN) classifier to extract their motion segments. Finally, a solution related to this work and combining the heuristic analysis with SVM classification has been proposed in [10].

Recognition: The CGC 2014 winning team [9] based their contribution on a multi-scale-DNN classifiers fusion operating at frame level. They learned from a combination of depth, RGB and a rich joint description. Similar DNN derived architectures were also used in [11, 12]. In these works, the depth and RGB features were automatically generated within the DNN.

Different works focused on the use of hand-crafted features in combination with SVM classifiers. A joint-only-solution based on the joint-quadruplet descriptor was proposed in [8]. Another solution focusing on the RGB video streams and using the state of the art winning dense descriptor [13] was proposed in [4]. The authors produced a 'super vector' that received a Gaussian mixture model representation and later a linear SVM classification. A tentative to exploit all the existing modalities was proposed by Liang *et al.* [5] through a motion-trail descriptor. Although the solution had a recognition performance of 92.8%, its Jaccard index measure has been lower [1, 5].

Finally, a variety of approaches were proposed for different purposes. Monnier *et al.* [14] had the advantage of presenting the fastest 2-hour-learning classifier based on a multi-scale boosting-oriented description. Other solutions focused on the use of random forests [15] and extremely randomised trees [9] for their baselines. A common pre-processing stage found in most approaches is related to the identification of the left/right dominant hand in motion.

Approach proposition: The analysis of the presented works shows that a small portion of them focused on the efficient combination of all of depth, mask, RGB and joint modalities together. In addition, none of the works focused on the use of the facial data. Theoretically, the more modalities we use, if combined using the right strategy, the better performances we get. We propose in this paper a complementary combination of all of the existing modalities. We also add the facial data to produce a richer hand-crafted feature description ready for SVM classification. The details of our solution represented in Fig. 1 are going to be discussed hereafter.

3. TEMPORAL SEGMENTATION

The TS approach used in this work is similar to the one we presented in [10] with improvements relative to the introduction of the pairwise distances, the feature selection and learning label configuration. Here we rapidly review the existing processing then detail the newly introduced ones.

Joint Augmented Context: As the joints convey sufficient information about the existence of motion, we rely on this unique modality for the TS purpose. Our interest goes into the 11 upper-body joints. As they come with inherent noise and variable sizes, we start by filtering noisy behaviours using a mean filter on the time (t) axis, then translate to the hip-center J^C position and normalise the skeletal lengths to the unit size using an automatically determined scaling factor (S). The stabilised and normalised 3D joints $\mathfrak{J}^i, i \in [1..11]$ produce our first 33 features (3x11) as in (1):

$$\mathfrak{J}^i_{x,y,z}(t) = (J^i_{x,y,z}(t) - J^C_{x,y,z}(t)) \times S_{x,y,z}(t) \tag{1}$$

The extracted features produce a shape-context-like augmented representation [16]. We start by concatenating the joint's pairwise distance vectors given by (2):

$$pwDist = \| \mathfrak{J}^i_{x,y,z} - \mathfrak{J}^j_{x,y,z} \| , \ \ with \ \ i \neq j \tag{2}$$

In addition, we recuperate the quaternion angles Q information offered within the CGC dataset and composed of 4 values $[q_w, q_x, q_y, q_z]$ per joint. Finally, we extract 66 features relative to temporal gradients δ and δ^2 of first and second order as in (3) and (4) respectively:

$$\delta^i(t) = (\mathfrak{J}^i_{x,y,z}(t+1) - \mathfrak{J}^i_{x,y,z}(t-1)) \tag{3}$$

$$\delta^{i2}(t) = \mathfrak{J}^i_{x,y,z}(t+2) - 2\mathfrak{J}^i_{x,y,z}(t) + \mathfrak{J}^i_{x,y,z}(t-2) \tag{4}$$

TS feature optimisation: The generated representation contains a feature vector of 264 values that encapsulate all the useful information about the pose and temporal motion development. In order to select the pertinent vectors, we have applied the rapid feature selection approach suggested by Vervidis *et al.* [17] over our CGC 2014 based features. It measures each feature's pertinence by computing the Correct Classification Rate (CCR) of a naive Bayesian classifier. The details about the Sequential Floating Forward Selection (SFFS) strategy employed can be found in [17]. The generated feature vector out of this stage contains 251 descriptors.

Ground truth label extension: Similar to the approaches [4] and [5], we compare the $\mathfrak{J}^{LH}$ and $\mathfrak{J}^{RH}$ left-and-right hand 2D joints positions towards an anchor point (the $\mathfrak{J}^C$ hip-centre joint position in our case) to find motion-segments. Our heuristic analysis conditioned by a threshold τ outputs:

$$\begin{cases} 1 \, if \ \ ((\mathfrak{J}^{LH}_{x,y} - \mathfrak{J}^C_{x,y}) > \tau \ \ and \ \ (\mathfrak{J}^{RH}_{x,y} - \mathfrak{J}^C_{x,y}) > \tau) \\ 0 \, elsewhere \end{cases}$$

The obtained motion segments are used only to enrich the ground truth motion label positions passed to our SVM linear classifier. This step ensures a better inter-class variability between the used features.

TS feature classification: Following the approach presented in [10], we have used a concatenation of 5 consecutive frame descriptors as a unique, per frame, feature vector. This configuration has allowed better action separation and robustness towards empty-joint frames. The obtained 1255-feature vector, coupled with the enriched ground truth binary labels obtained fro the previous step, allowed the learning of a linear SVM classifier. Its classification frame-wise decisions produced our TS through motion/non-motion distinction.

4. ACTIONS RECOGNITION AND GROUPING

We present in this section the different features used, then we detail our frame labelling and gesture recognition strategy.

4.1. Feature multi-modal representation

The features used for the action recognition stage combine the joint descriptor presented in the previous section with the features extracted from the RGB, mask and depth video streams. Compared to similar works related to the CGC context, we extract our features from the hand regions but we also add additional descriptors relative to the face region. An overview of the used features is presented in Fig.2. In what follows, we give the details of the extracted additional learning data.

Colour HOG (CHOG): From the RGB video streams, we extract the HOG features relative to both left and right hands in addition to the face. The bounding boxes of 64x64 are extracted from the stabilised 2D joint positions and the left hand image is flipped to make it similar in aspect with the right one. This step is useful for a later dominant hand analysis. Each bounding box is rescaled to 48x48 from which 8 gradient orientations are extracted using 4 sub-cells. This operation produces 3 vectors where each one has 32 values.

Depth Motion Histograms (DMoHist): For the depth modality, we first subtract the background by multiplying the depth pixels with those of the available mask, then we compute the motion difference $\Delta(t)$ between any $(t+1)$ and $(t-1)$ sets of frames. Using the stabilised 2D joints positions, we extract 3 bounding boxes of 64x64 relative to the hands and face. After flipping the left hand, each region is downscaled to a 4x4-resolution grid. This process generates a 48-bin histogram of depth differences relative to our DMoHist descriptor given by (5):

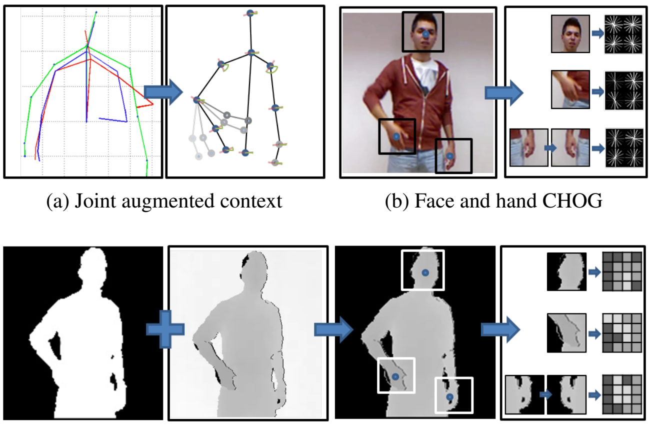
$$\Delta_i(t) = (d_i(t+1)\times m_i(t+1)) - (d_i(t-1)\times m_i(t-1)) \quad (5)$$

where $i \in [1..16]$ is the downsized bloc index, $d(t)$ and $m(t)$ are respectively the depth and mask of frame t.

Feature representation: For both depth and RGB hand-relative features, we apply an additional processing step related to the identification of the dominant hand in motion. As a given action can be performed by any of the hands, we ensure that dominant hand features are placed first in the feature vector passed to classification. This allows the classifier to easily and robustly find decision boundaries. For this purpose, we compute, for every action of N frames, the 3D cumulated trajectories of both left and right hand joints using (6):

$$D(\mathfrak{J}^i_{x,y,z}) = \sum_{1}^{N-1} |\mathfrak{J}^i_{x,y,z}(t+1) - \mathfrak{J}^i_{x,y,z}(t)| \quad (6)$$

In case $D(\mathfrak{J}^{LH}_{x,y,z}) > D(\mathfrak{J}^{RH}_{x,y,z})$, we swap the features' order. The analysis of the feature pertinence, using the SFFS solution presented in section 3, has proved the necessity of all features in disposal for better classification performances.



(a) Joint augmented context

(b) Face and hand CHOG

(c) Face and hand masked-depth motion histograms

Fig. 2: Feature extracted from the different input modalities

4.2. Label classification and grouping strategy

The goal of this process is to generate unique labels for the extracted TS regions by regrouping single-frame labels.

Label learning: During the learning stage, we have used an updated version of the groundd truth. We have generated an augmented sequence of frame labels containing, in addition to the labels 1 to 20 relative to the vocabulary, label 0 for non-motion and label 21 for non-vocabulary actions to make a total of 22 labels. This labelling entry has allowed us to further improve the segmentation using the generated 0 labels. Furthermore, any action segment identified with the extra label 21 has been automatically rejected.

For label recognition, we have trained a linear SVM classifier using a concatenation of 251 joint-based features with 96 CHOGs and 48 DMoHist ones. All the obtained 395 features have been synchronised and normalised to zero mean and unit variance.

Action extraction and grouping: In order to produce unique action labels for any given sequence of extracted frames, we have permitted frame-wise action recognition, and then applied label grouping and extraction. This choice has enabled our classifier to learn from a richer descriptor population compared to the bag of words representation.

As presented in Algorithm 1, we have extracted, for every action sequence s, the different actions using our TS approach and generated their frame-wise labels. If a given action was larger than a threshold $nf1$ (determined from the exhaustive analysis of all learning-action lengths), we further segmented it into coherent consecutive label segments.

As a given action presents variable labels at its borders (starting and ending at rest positions), we have extended the obtained segments by $nf2$ frames. The final decision about the global label for all considered action-frames has been determined by the most occurring label within the middle frames delimited by $nf3$ as detailed in Algorithm 1.

Algorithm 1: Segment labelling strategy

```
Input:  $L_{s,a}$ labels for sequence $s$ and action $a$
        $TS_{s,a}$ action border couples $(start_a, end_a)$
Output: $\mathfrak{L}_s$ global labels within the sequence $s$
for each test sequence $s$ do
    Extract action list $a$ borders
    for each action $a$ do
        if (length($L_{s,a}$) $> nf1$) then
            // too long action
            Extract consecutive labels sub-actions $a_{sub}$
            for each sub-action $a_{sub}$ do
                $nf2 \leftarrow round(length(a_{sub})/20)$
                Enlarge $a_{sub}$ by $nf2$ frames
            Update action list $a$ borders and labels
    for all new actions $a$ do
        //Action global-label decision
        $nf3 \leftarrow round(length(L_{s,a})/6)$
        Extract middle sub-labels $L_{sub}$ delimited by $nf3$
        Add most occurring label inside $L_{sub}$ to $\mathfrak{L}_s$
```

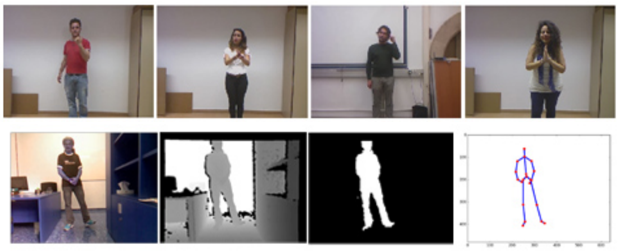
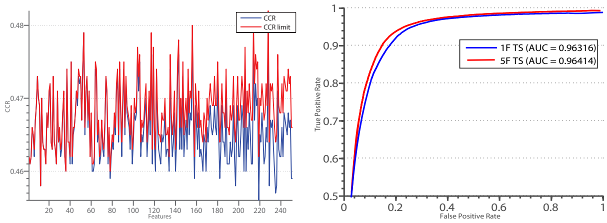


Fig. 3: The first row shows samples of actors and the actions *viene_qui*, *messi_daccordo*, *sei_furbo* and *cosa_ai_combinato* respectively. The second row illustrates consecutively the color, depth, mask and joints modalities in our dataset.



(a) Features Selection CCR

(b) TS ROC

Fig. 4: a- The CCR improvement using the naive Bayesian classifier for feature selection. b- The 5-frame TS classifier presents a useful ROC area under curve improvement compared to the 1-frame base one.

5. EXPERIMENTS

We present in what follows the configuration set and performance of the separate processing stages, then their combined evaluation using the mean Jaccard index.

CGC dataset: The CGC 2014 is a multi-actor public dataset containing motion streams captured using kinect sensor for 1 person at a time. Each actor is asked to arbitrarily realise 20 kinds of sign language actions relative to the Italian vocabulary. The performances are hand based with resting poses, but the streams may contain challenging unknown or consecutive un-separated ones. The CGC 2014 offers 3 sub-datasets used for learning, validation and test evaluation. Each one contains data streams relative to the joints, depth, mask and RGB in addition to the ground truth label files. Figure 3 gives an idea about the CGC 2014 dataset contents.

Temporal segmentation performance: We show in Fig.4a the development of the CCR using the SFFS feature selection criterion. This step accelerated our models learning time and improved their performance with small percentiles.
Using 1 frame or 5 frames feature configurations, we obtained similar classification performances near 92.13% and an area under curve near 0.964. But, the second configuration has been adopted as is allowed a more separative behaviour, specially in continuous action cases. The observed performance for the binary SVM classifier is reported in Fig.4b.

Action classifier performance: The evaluation of our classifier trained on 22 labels using enabled us to obtain an average recognition rate of 81.01%. 1 shows example CCR results for a selection of video sequences relative to the test set. It shows the contribution of the grouping stage to the overall recognition quality. The comparison with other classifiers learned using different sets of 20 or 21 labels proved the superiority of the 22-label SVM model. In Tab.2, we present a reduced representation of the confusion matrix relative to our used classifier.

Approach evaluation: To evaluate our solution, we used the Jaccard index measure as presented in [1]. It permits the evaluation of both action positions and label exactitude as shown in (7):

$$Jaccard(A,B)_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}} \qquad (7)$$

where $A_{s,n}$ is the obtained action with name n for a sequence s and $B_{s,n}$ is the ground truth. Using this measure, the evaluation of classifier without facial data generates a Jaccard distance of 0.5667. After adding the facial data extracted from the depth and RGB modalities, the Jaccard distance obtained reaches 0.6180. These performances prove the effectiveness of our assumption concerning the importance of the facial data for action recognition.
The analysis of the separate label performances showed that our extra descriptors where mainly useful for the 15 CGC one-handed actions where there is more contact between the hand and the face region. Thus, our richer description allowed a better SVM classification of those actions.
Compared to the state of the art, and though we have not used deep neural learning architectures, our solution still proves its competitiveness and can be further improved using multiple

modalities/body-parts classifiers fusion. Table 3 positions our overall performance within the state of the art.

6. CONCLUSION

We have presented in this paper an approach that allows the extraction and recognition of actions from continuous streams of multi-modal inputs. Our solution first applies a temporal segmentation using the SVM classification of an augmented joints data representation. Afterwards, it gradually adds features relative to the RGB and depth modalities with an awareness to the dominant hand and face regions. It classifies frame-by-frame the labels then applies a grouping strategy in order to produce coherent segment labels.
Compared to the state of the art, we have proved the contribution of the facial data and obtained a competitive ranking using the Jaccard index measure. Our method could be further improved by applying multi-modal and multi-scale decisions fusion through multiple classifier families. These potential perspectives are already in progress.

	Default Recog.	**Improved Recog.**	**Jaccard Distance**
Sample0770	55,55%	69,36%	0,60005
Sample0777	70,32%	98,73%	0,95827
Sample0840	43,15%	45,85%	0,32326
Sample0880	55,55%	64,06%	0,71297
Sample0908	67,30%	100%	0,82487
Sample0927	59,94%	86,91%	0,81150

Table 1: Classifier evaluation on a subset of the test dataset. The evaluation of the obtained labels is generated using the ground truth augmented labels for every TS segment.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.84	0.7	0.58	0.87	0.88	0.87	0.92	0.84	0.98	0.62

11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0.63	0.83	0.95	0.7	0.53	0.96	0.86	0.75	0.86	0.9

Table 2: The 20 separate-label recognition rates generated from the confusion matrix. The classifier learned from a concatenation of all modalities has an accuracy of 81.01%

Proposed solution	**Jaccard distance**
Neverova *et al.* [9]	0.8500
Evangelidis *et al.* [8]	0.7454
Our Approach	**0.6180**
Liang *et al.* [5]	0.5971
Team 'YNL' [1]	0.2706

Table 3: Our solution ranks in a competitive position amongst existing literature using the Jaccard distance mesure on CGC 2014 [1].

REFERENCES

[1] S. Escalera, X. Bar, J. Gonzlez, M.A. Bautista, M. Madadi, M. Reyes, V. Ponce-Lpez, H.J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *ECCV Workshops*, Sept. 2014.

[2] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. Escalante, "The chalearn gesture dataset (cgd 2011)," *Mach. Vis. and Applications*, vol. 25, no. 8, pp. 1929–1951, 2014.

[3] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos, "Multimodal gesture recognition via multiple hypotheses rescoring," *Journal of Machine Learning Research*, vol. 16, pp. 255–284, 2015.

[4] X. Peng, L. Wang, and Z. Cai, "Action and gesture temporal spotting with super vector representation," in *ECCV Workshops*, Sept. 2014.

[5] B. Liang and L. Zheng, "Multi-modal gesture recognition using skeletal joints and motion trail model," in *ECCV Workshops*, Sept. 2014.

[6] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IPAMI*, vol. 31, no. 9, pp. 1685–1699, 2009.

[7] J. Y. Chang, "Nonparametric Gesture Labeling from Multi-modal Data," in *ECCV Workshops*, Sept. 2014.

[8] G. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in *ECCV Workshops*, Sept. 2014.

[9] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *ECCV Workshops*, Sept. 2014.

[10] B. Seddik, S. Gazzah, T. Chateau, and N. Essoukri Ben Amara, "Augmented skeletal joints for temporal segmentation of sign language actions," in *IEEE IPAS'14*, Hammamet, Tunisia, Nov. 2014.

[11] L. Pigou, S. Dieleman, and P.-J. Kindermans, "Sign Language Recognition Using Convolutional Neural Networks," in *ECCV Workshops*, Sept. 2014.

[12] D. Wu, "Deep Dynamic Neural Networks for Gesture Segmentation and Recognition," in *ECCV Workshops*, Sept. 2014.

[13] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, 2013.

[14] C. Monnier, S. German, and A. Ost, "A Multi-scale Boosted Detector for Efficient and Robust Gesture Recognition," in *ECCV Workshops*, Sept. 2014.

[15] N. C. Camgoz, A. A. Kindiroglu, and L. Akarun., "Gesture Recognition using Template Based Random Forest Classifiers," in *ECCV Workshops*, Sept. 2014.

[16] S. Belongie, J. Malik, and J. Puzicha, "Matching shapes," in *ICCV*, July 2001, vol. 1, pp. 454–461.

[17] D. Ververidis and C. Kotropoulos, "Information loss of the mahalanobis distance in high dimensions: Application to feature selection," *TPAMI*, vol. 31, no. 12, pp. 2275–2281, Dec 2009.