Folk Dance Recognition using a Bag of Words Approach and ISA/STIP Features

Ioannis Kapsouras, Stylianos Karanikolos, Nikolaos Nikolaidis and Anastasios Tefas
Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 541 24, GREECE
{nikolaid,tefas@aiia.csd.auth.gr}

ABSTRACT

Recognition of folk dances i.e. classification of dance videos according to the specific dance depicted can be considered a challenging sub task within the general activity recognition area because of the large number of different dances, the similarities among them and the different styles a dance can be performed. A method able to identify various folk dances is very important for analyzing and annotating multimedia databases of such dances thus helping the preservation of folk dance culture. In this paper, we deal with recognition of Greek folk dances. Clustering is applied on input features to extract a codedbook and a bag of words approach is applied. An SVM classifier is used for the classification. Two state of the art methods for feature extraction are used and compared. The method is applied on two folk dances from the Western Macedonia region.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motion*; I.4 [**Image Processing and Computer Vision**]: Scene Analysis

1. INTRODUCTION

Activity recognition is an active research topic that deals with the process of labeling a motion sequence with respect to the depicted motions. Activity recognition is very important for various applications such as video surveillance, video annotation, human computer interaction etc. Most of the proposed algorithms for activity recognition deal with everyday activities such as walking, running, jumping etc. A great amount of research has been performed on activity recognition. A survey of activity recognition approaches can be found in [7], [9], [2]. Many methods of activity recognition use global representations whilst others extract features from local areas. Classification can be performed by many ways, such as nearest neighbour, SVM, HMM, dynamic time wrapping etc. However dancing is a very wide motion class and thus recognition of dances i.e. classification of dance videos according to the depicted dance, is considered as a different research field. There are a lot of dances that can be recognized as different activities such as tango, breakdance, waltz etc. Although video based activity recognition is a very active research field, research on dance recognition is very limited. Samanta et al. in [8] propose a method for classifying Indian Classic Dances. The authors propose a pose descriptor to represent each frame of a sequence. The descriptor is based on histogram of oriented optical flow, in a hierarchical manner. The pose basis is learned using an on-line dictionary learning technique and each video is represented sparsely as a dance descriptor by pooling pose descriptor of all the frames. Finally, dance videos are classified using support vector machine (SVM) with intersection kernel. Two methods for dance pose recognition (which is a task related to dance recognition) are presented in [6], [1].

In this paper we use a bag of words approach to perform folk dance recognition. K-means is applied on features extracted from the training data to create a codebook. Then vector quantization is applied and a histogram over the code words for each sequence is created. Finally, an SVM classifier with χ^2 kernel is applied on the histograms for classificaiton. For feature extraction, two state of the art methods that generate features suitable for activity recognition were used and compared. The first one, proposed by Le et. al [5], extends the Independent Subspace Analysis algorithm to learn spatio-temporal features from video data. The second, proposed by Laptev et. al [4], detects spatio-temporal interest points using an extension of the Harris detector (Harris3D).

2. PROBLEM STATEMENT

Folk music and folk dances constitute a significant part of the folk heritage around the world. The preservation of the folk music and choreographies and their dissemination to the younger generations is a very important issue since folk dances form an important part of a country's or region's history and culture.

Greece has a great tradition of folk dances with different rhythms and dancing styles in different regions of the country. A great variety of dances exists even within a specific region. The recordings of such dances are often of low quality and with no annotation. There are many dancing groups performing various dances in various festivals however cataloging and recording of folk dances e.g. in a central annotated video database of folk dances is very scarce. Moreover, some folk dances are known only to some senior citizens and the danger of their choreographies being forgot-

BCI '13 , September 19-21, 2013, Thessaloniki, Greece.

ten is existent. An annotated folk dances database will be of great importance for educational, research and cultural heritage preservation purposes. Such a database will help the youngsters to stay in touch with their cultural heritage and increase their awareness for it.

A system that can perform dance recognition from a video is very important for the creation and the annotation of a multimedia database. Folk dances recognition and classic activity recognition bear important similarities such as the need of robust feature extraction and the selection of a good classifier, but the two tasks have also differences mainly in the input data. First of all, Greek folk dances are mostly group, circular dances so a recognition method has to take into account that there are multiple subjects (dancers) in a video. Furthermore, the long skirts of the women dancers hide the legs making the recognition more difficult as shown in Figure 1. In addition, the traditional costumes may differ from place to place and between various dancing groups making activity recognition methods that rely on appearance less effective. Also many folk dances have similar steps and tempo, so the recognition between them is a challenging task. In addition, in certain geographical areas such as Western Macedonia, the tempo of some folk dances changes from slow to fast with, sometimes, a change in the steps of the dance, making the recognition even harder because of the significant inter-class variation. Moreover, the dances can take place either indoors or outdoors as shown in Figure 2. Finally recognizing folk dances on recordings from folk festivals and fairs is very difficult since a lot of amateur dancers participate and the scene is often very crowded as can be seen in Figure 3. In summary, folk dance recognition can be considered as a more challenging research field than general activity recognition.

Figure 1: Women dancers wearing long skirts.

3. METHOD DESCRIPTION

The aim of this paper is to apply a recognition framework used in general activity recognition on the specific sub task of folk dance recognition. More specifically, we aim to check whether such an approach can operate sufficiently on folk dance videos. Moreover a comparison of the performance of two different state of the art feature extraction approaches are presented.

Feature vectors are used to represent the video data. K-means is applied on feature vectors to create a codebook consisting of the centroids $\mathbf{v}_c, c = 1, \cdots, C$ where C is the number of the K-means clusters. Each centroid represents a code word of the codebook. Then, the feature vectors are mapped to its closest code word using *Euclidian* distance. Next for each training sequence the frequency of appearance of every codeword is computed and thus, a histogram for each sequence that characterizes is formed.

Figure 2: Greek folk dances performed by professional dancing groups: a) Lotzia dance performed indoors, b) Lotzia dance performed outdoors.

Figure 3: Greek folk dance performed by many amateur dancers in a festival.

For an unlabeled dance video sequence the same procedure is used, thus, the feature vectors of the test sequence are mapped to the closest codewords of the codebook and the histogram of codewords that characterizes the test sequence is formed. Then, an SVM classifier trained using the histograms of the training set is used for the classification. We used a non-linear SVM with χ^2-kernel [3]:

$$K(\mathbf{s}_j, \mathbf{q}_{test}) = exp(-\frac{1}{2A}\sum_{i=1}^{C}\frac{(s_{j,i} - q_{test,i})^2}{s_{j,i} + q_{test,i}}) \qquad (1)$$

where C is the codebook size and $s_{j,i}$ and $q_{test,i}$ are the values of the i-th bin for the histogram $\mathbf{s}_j$ of the j-th training sequence and the test sequence histogram $\mathbf{q}_{test}$, respectively. The above method is a bag of words approach proposed in [10] for evaluating features for activity recognition. We used this method to compare two types of features proposed in [5] and [4] in folk dances recognition. These features are described in the following.

Le et al. use Independent Space Analysis (ISA) to learn unsupervised features from a video. An ISA network can be described as a two-layered network with square and square-root nonlinearities in the first and second layer respectively. In more detail:

$$p_i(\mathbf{x}^t; \mathbf{W}, \mathbf{V}) = \sqrt{\sum_{k=1}^{m} V_{ik} (\sum_{j=1}^{m} W_{kj} x_j^t)^2} \quad (2)$$

is the activation of each second layer unit given an input pattern $\mathbf{x}^t$. Parameters $\mathbf{W}$ are learned through sparse representation in the second layer by solving:

$$\begin{aligned} &\underset{\mathbf{W}}{\text{maximize}} \sum_{t=1}^{T} \sum_{i=1}^{m} p_i(\mathbf{x}^t; \mathbf{W}, \mathbf{V}) \\ &\textit{subject to } \mathbf{W}\mathbf{W}^T = \mathbf{I} \end{aligned} \quad (3)$$

where $\mathbf{W} \in \Re^{k \times n}$ is the matrix that contains the weights connecting the input data to first layer units and $\mathbf{V} \in \Re^{m \times k}$ is the matrix that contains the weights connecting the units of the first layer to second layer units.

Le et al. use $3D$ video blocks (patches) as input to the first layer of the neural network. In order to reduce the cost of the algorithm, they use small patches and convolve the trained network by overlapping the first layer trained features to compute the input of the second layer of the network. PCA is used as a preprocessing step to reduce the dimension of the input data. Finally, they combine features from both layers and use them for classification. Their method is trained using a batch projected gradient descent. Examples of first layer's learned features are shown in Figure 4.

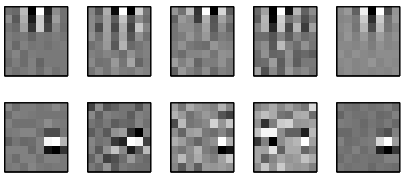

Figure 4: Examples of two ISA features learned from folk dances data (size of the video block: $8 \times 8 \times 5$).

Laptev et al. in [4] propose a method that is based on the evaluation of Space-Time Interest Points (STIPS) from each action video and their description by a set of Histograms of Oriented Gradients/Histograms of Optical Flow (HOG/HOF) descriptors, which refer to local shape and motion. The authors employ the Harris3D detector, which was proposed by Laptev and Lindeberg in [3], in order to detect video locations where the image intensity values undergo significant spatio-temporal changes. Harris3D extends the Harris interest point detector and the basic idea is to extend the notion of interest points in the spatial domain, by requiring the image values in local spatio-temporal volumes to have large variations along both spatial and temporal directions. Points with such properties are STIPS and they correspond to local spatio-temporal neighbourhoods with non-constant motion. The authors construct the linear scale-space representation of spatio-temporal image sequence f, by convolution of f with an anisotropic Gaussian kernel, with independent spatial and temporal scale values σ_l^2 and τ_l^2.

$$L = (\cdot; \sigma_l^2, \tau_l^2) = g(\cdot; \sigma_l^2, \tau_l^2) * f(\cdot) \quad (4)$$

Then, they consider a spatio-temporal second-moment matrix $\mathbf{M}$ at each video point, which is a 3×3 matrix composed of first order spatial and temporal derivatives averaged with a Gaussian function $g(\cdot; \sigma_i^2, \tau_i^2)$, with $\sigma_i^2 = s\sigma_l^2$ and $\tau_i^2 = s\tau_l^2$. The final locations of space-time interest points are given by local maxima of $H = det(\mathbf{M}) - ktrace^3(\mathbf{M})$, $H > 0$. The HOG/HOF descriptors are used to compute histograms of oriented gradient and optical flow, accumulated in space-time volumes in the neighbourhood of detected interest points. The size of each volume is related to the detection scales by Δ_x, $\Delta_y = 2k\sigma$ and $\Delta_t = 2k\tau$. Each volume is subdivided into a $n_x \times n_y \times n_t$ grid of cuboids and for each cuboid, histograms of gradient orientations and histogram of optical flow are computed. Finally, normalized histograms are concatenated into HoG, HoF as well as HoG/HoF descriptor vectors. Sample STIPs detected in a folk dance video are shown in Figure 5.

(a) (b)

Figure 5: Spatio temporal interest points detected in a folk dance video: a) Lotzia dance performed outdoors, b) Capetan Loukas dance performed outdoors.

4. EXPERIMENTAL RESULTS

The features proposed in [5] and [4] were tested in standard datasets for classic activity recognition. We have tested these features within the recognition framework presented in the previous section on a small set of videos of Greek folk dances in order to verify the ability of these feature extraction methods to cope with the particularities of folk dance recognition. The dataset contains 4 videos of 2 Greek folk dances, namely *Lotzia* and *Capetan Loukas*. More precisely, 1 video from *Lotzia* and 1 video from *Capetan Loukas* (Figure 2), both performed by professional dancing groups indoors were used for training and the other 2 videos, performed outdoors, were used for testing. The difference in recording conditions (indoor/outdoor) between the training and testing videos makes the recognition setup more realistic but also more challenging. The indoor videos of *Lotzia* and *Capetan Loukas* were temporally segmented into overlapping clips of duration 80 to 100 frames each resulting to 78 and 113 clips respectively. The same procedure was used for the outdoors videos resulting to 102 and 107 clips respectively. Thus the training and test set consisted of 191 and 209 clips respectively. The two methods were tested for various numbers of K-means clusters C and the best results are presented in Table 1.

Table 1: Comparison of classification performance on the folk dances dataset.

Method	Classification Rate	Number of clusters
ISA [5]	89.47%	2000
STIP [4]	87.08%	30

As can be seen in Table 1 features extracted from ISA method outperform the STIPS features by almost 2%. Despite the fact that there are only two different dance classes the importance of achieving a sufficiently high recognition rate should not be underestimated since as mentioned in Section 2 the task is a very challenging one. It is obvious from Table 1 that the two methods achieve the best classification rate for very different number of clusters. This can be attributed to the fact that STIP extracts feature vectors only for interest points whereas ISA extracts such vectors in a global manner.

5. CONCLUSIONS

In this paper, we deal with recognition of Greek folk dances. Clustering is applied on input features to extract a codedbook and a bag of words approach is applied. An SVM classifier is used for the classification. Two state of the art methods for feature extraction are used and compared. The method is applied on two folk dances from the Western Macedonia region. The results are very promising. Fusion of the proposed features and experiments on much larger datasets will be topics for future research.

6. ACKNOWLEDGMENTS

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operation Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS-UOA-ERASITECHNIS MIS 375435.

7. REFERENCES

[1] X. Feng and P. Perona. Human action recognition by sequence of movelet codewords. In *First International Symposium on 3D Data Processing Visualization and Transmission, 2002. Proceedings.*, pages 717–721, 2002.

[2] X. Ji and H. Liu. Advances in view-invariant human motion analysis: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(1):13–24, 2010.

[3] I. Laptev and T. Lindeberg. Space-time interest points. In *Ninth IEEE International Conference on Computer Vision, 2003. Proceedings.*, pages 432–439 vol.1, 2003.

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

[5] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011*, pages 3361–3368, 2011.

[6] B. Peng and G. Qian. Binocular dance pose recognition and body orientation estimation via multilinear analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08.*, pages 1–8, 2008.

[7] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, June 2010.

[8] S. Samanta, P. Purkait, and B. Chanda. Indian classical dance classification by learning dance pose bases. In *IEEE Workshop on Applications of Computer Vision (WACV), 2012*, pages 265–270, 2012.

[9] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[10] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, page 127, sep 2009.