Feature Comparison and Feature Fusion for Traditional Dances Recognition

Ioannis Kapsouras, Stylianos Karanikolos, Nikolaos Nikolaidis,
and Anastasios Tefas

Department of Informatics,
Aristotle University of Thessaloniki,
54124 Thessaloniki, Greece
{jkapsouras,nikolaid,tefas}@aiia.csd.auth.gr

Abstract. Traditional dances constitute a significant part of the cultural heritage around the world. The great variety of traditional dances along with the complexity of some dances increases the difficulty of identifying such dances, thus making the traditional dance recognition a challenging subset within the general field of activity recognition. In this paper, three types of features are extracted to represent traditional dance video sequences and a bag of words approach is used to perform activity recognition in a dataset that consist of Greek traditional dances. Each type of features is compared in a stand alone manner in terms of recognition accuracy whereas a fusion approach is also investigated. Features extracted through the training of a neural network as well as fusion of all three types of features achieved the highest classification rate.

Keywords: Dance recognition, Dense Trajectories, Spatio-temporal interest points, Subspace Analysis, Neural networks.

1 Introduction

Activity recognition is an active research topic that deals with the identification of activities performed by a human subject as captured in video. Activity recognition deals mainly with the recognition of everyday actions such as walking, running, sitting etc. and is important for various application such as video surveillance and video annotation. Moreover, activity recognition can be used for creating intelligent environments, for computer-human interaction etc. A significant amount of research has been performed on activity recognition. Surveys of activity recognition approaches can be found in [1], [2], [3]. 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. Methods for activity recognition can also work on multi view sequences.

Dancing is a very wide motion class that includes many different syles (e.g. tango, breakdance, waltz, traditional dances etc) and has many particularities. Thus recognition of dances can be considered as a different research field.

L. Iliadis, H. Papadopoulos, and C. Jayne (Eds.): EANN 2013, Part I, CCIS 383, pp. 172–181, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Although video based activity recognition is a very active research field, research on dance recognition both on video and motion capture data is very limited. Samanta et al. in [4] 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 the 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 descriptors of all the frames. Finally, dance videos are classified using support vector machine (SVM) with intersection kernel. In [5] Raptis et al. introduce a method for real-time classification of dance gestures from skeletal animation. An angular skeleton representation that maps the motion data to a smaller set of features is used. The full torso is fitted with a single reference frame. This frame is used to parametrize the orientation estimates of both the first-degree limb joints (joints adjacent to torso) and second-degree limb joints (tips of the wireframe extremities such as the hands and the feet). Then a cascaded correlation-based maximum-likelihood multivariate classifier is used to build a statistical model for each gesture class. The trained classifier compares the input data with the gesture model of each class and outputs a maximum likelihood score. An input gesture is finally compared with a prototype one using a distance metric that involves dynamic time-warping. Deng et al. in [6] proposed a method that performs recognition of dance movements on skeletal animation data. The authors proposed a new scheme for motion representation, the segmental SVD. A motion pattern is represented in a hierarchical structure with multiple levels and SVDs generated on the corresponding levels are used to extract features across time. The authors also proposed a similarity measure to compare segmental SVDs representations. Two methods for dance pose recognition (which is a task related to dance recognition) are presented in [7], [8].

In this paper we use a bag of words approach to perform traditional dance recognition on video data. Features extracted from the training data are clustered using K-means to find discriminative representations of the features. Then each feature vector is mapped to the closest cluster center and a histogram over the cluster centers is created. An SVM classifier with χ^2 kernel is trained to classify an unknown sequence of a traditional dance. Three state of the art methods used in general activity recognition research were used for feature extraction. The first one, proposed by Le et. al [9] extends the Independent Subspace Analysis algorithm to learn spatio-temporal features from video data by training a neural network. The second, proposed by Laptev et. al [10] detects spatio-temporal interest points using an extension of the Harris detector (Harris3D). The third, proposed by Wang et al [11] represents a video sequence based on dense trajectories and motion boundary descriptors. The performance of each individual type of features in the recognition of 5 Greek traditional dances is experimentally evaluated. Furthermore, two fusion approaches are investigated and compared.

2 Problem Statement

Parts of history of the world and various traditional customs are reflected in traditional music and dances. There is a great variation of traditional dances and the preservation and dissemination of such dances to the younger generations is a very important issue for a specific country or region.

In Greece, there are many traditional dances due to its rich history and cultural diversity. There is a great variation between dances even within a specific region. There are fast and slow dances, dances performed only by women and dances that change tempo from slow to fast. The recordings of such dances are usually of low quality and with no annotation. Some dances are very rare and known only to some senior citizens. An annotated traditional 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.

Traditional dance recognition can be considered more challenging than generic activity recognition. A system that can recognize traditional dances needs a robust feature selection procedure and a reliable classifier, both of which are also parts of a general activity recognition algorithm. However traditional dance recognition has important particularities and difficulties. At first, there are traditional dances that have similar tempo and steps making the recognition between them more difficult. Moreover, the rhythm of some songs changes from slow to fast within the song thus affecting the tempo of the dance. These changes increase the inter-class variation and thus the difficulty of recognition. Furthermore, most of the Greek traditional dances are group, circular dances and it is highly unlikely that methods designed for activity recognition in one subject will achieve high recognition rates when dealing with many subjects (dancers). Finally, professional dancing groups often perform the same dance by traditional costumes that differ as shown in Fig. 1 making activity recognition methods that rely on appearance less effective.

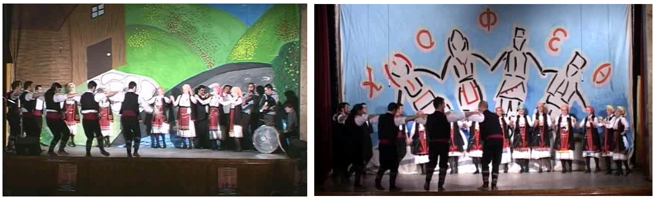

Fig. 1. Stankena Greek folk dance performed by professional dancing groups with different costumes

3 Method Description

The aim of this paper is to test if a well-known framework applied in general activity recognition can be used with good results to recognize Greek traditional dances. To test this framework three different state of the art feature extraction approaches are presented and compared. Moreover, the fusion of these features in two different ways has been considered.

The bag of words recognition framework [12] is summarized as follows. Let a number of feature vectors represent the video data. In order to recognize a number of dance classes, feature vectors of training data are clustered using the K-means algorithm. The centroids $\mathbf{v}_c, c = 1, \cdots, C$ where C is the number of clusters of the K-means, form a discriminative representation of the feature vectors. Then the feature vectors of all the training data are mapped to the closest centroid using *Euclidean* distance. Next for each training sequence the frequency of appearance of every centroid is computed and thus, a histogram for each sequence, that characterizes it, is formed.

Feature vectors are also extracted for a testing sequence and the same procedure is used. Thus, the feature vectors are mapped to the closest centroid and the histogram that characterizes the testing sequence is formed. At last, the testing sequence is recognized using an SVM classifier trained by the histograms of the training sequences. We used a non-linear SVM with χ^2 kernel [10]:

$$K(\mathbf{s}_j, \mathbf{s}_k) = exp(-\frac{1}{2A}\sum_{i=1}^{C}\frac{(s_{j,i} - s_{k,i})^2}{s_{j,i} + s_{k,i}}) \tag{1}$$

where A is the mean value of distances between all training samples, C is the number of centroids and $s_{j,i}$ and $s_{k,i}$ are the values of the i-th bin for the histograms $\mathbf{s}_j$ and $\mathbf{s}_k$. The *one-against-rest* approach was used for the SVM. As already mentioned, we used this framework to test three types of features proposed in [9], [10] and [11] in traditional dances recognition. Fusion of these features is also considered. The three features are described below.

Le et al. use unsupervised feature learning by training a neural network as a way to extract features from video data. The authors extend the algorithm of Independent Subspace Analysis (ISA). An ISA network can be described as a two-layered neural network with square and square-root nonlinearities in the first and second layer respectively. In more detail, the activation of each second layer unit for an input pattern $\mathbf{x}^t$ is given by:

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

Parameters $\mathbf{W}$ are learned through sparse representation in the second layer by solving:

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

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.

The authors use $3D$ video blocks (patches) as input to the first layer of the neural network. In order to reduce the computational 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 network is trained using a batch projected gradient descent. In what follows, the features generated by this approach will be denoted as ISA features.

Laptev et al. in [13] proposed a method for the determination of Space-Time Interest Points (STIPS) from each action video and their description by a set of histograms of oriented gradient (HOG) and histograms of optic flow (HOF) descriptors, which refer to local shape and motion. The authors employ the Harris3D detector, which was proposed by Laptev and Lindeberg in [10], 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 spatiotemporal 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 named STIPS and they correspond to local spatio-temporal neighbourhoods with non-constant motion. The authors construct the linear scale-space representation of a 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.

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

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 Fig. 2.

Wang et al. at [11] propose a method for activity recognition based on trajectories extracted by dense sampling. At first, dense sampling is performed on a grid spaced by W pixels. Sampling is performed in a number of spatial scales in order to track the sampled points through the video. In order to avoid samples in homogeneous image areas they use the criterion presented in [14] to remove points from these areas.

Feature points are tracked on each spatial scale separately by computing the optical flow field $\omega_t = (u_t, v_t)$ for each frame $\mathbf{I}_t$, where u_t and v_t are the

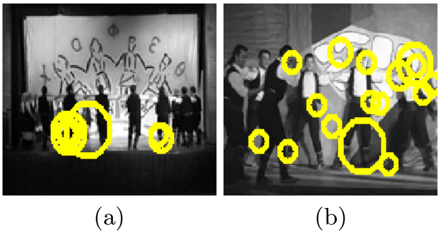

Fig. 2. Spatiotemporal interest points detected in folk dance videos: a) Stankena dance performed indoors, b) Zablitsena dance performed indoors

horizontal and vertical components of the optical flow. Given a point $\mathbf{P}_t = (x_t, y_t)$ in frame $\mathbf{I}_t$, its tracked position in frame $\mathbf{I}_{t+1}$ is smoothed by applying a median filter on ω_t. Points of subsequent frames are concatenated to form trajectories: $(\mathbf{P}_t, \mathbf{P}_{t+1}, \mathbf{P}_{t+2}, \ldots)$. The authors limit the length of trajectories to L frames. The shape of a trajectory is described by a sequence $(\Delta\mathbf{P}_t, \ldots, \Delta\mathbf{P}_{t+L-1})$ of displacement vectors, where $\Delta\mathbf{P}_t = (\mathbf{P}_{t+1} - \mathbf{P}_t)$. The resulting vector is normalized by the sum of displacement vector magnitudes. Trajectories extracted for a random frame from Stankena dance are shown in Fig. 3.

The authors also use a space-time volume aligned with a trajectory to encode motion information. The size of the volume is $N \times N$ pixels and L frames long and is subdivided into a spatio-temporal grid. In each cell of the spatio-temporal grid of the volume various descriptors are computed. These include HOG and HOF descriptors and motion boundary histograms (MBH) descriptors [15] in order to deal with camera motion. The final descriptor (that will be denoted as TRAJ) is computed via the concatenation of these descriptors.

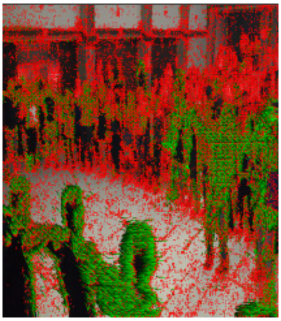

Fig. 3. Visualization of dense trajectories detected in a Stankena dance video

The above features were used in a bag of words manner, as described in Section 3, to train an SVM classifier with χ^2 kernel. Fusion of the above features was also considered. Fusion was performed either by adding the kernels of the SVM produced by the histograms of each type of features or by concatenating the histograms produced by different types of features.

4 Experimental Results and Discussion

The three methods for feature extraction described in the previous Section were used in order to verify if these features can successfully be used for the task of traditional dances recognition either individually or within a fusion framework. The features were tested within the framework presented in the Section 3 on a dataset of videos of Greek traditional dances. The dataset contains 10 videos of 5 Greek traditional dances, namely *Lotzia*, *Capetan Loukas*, *Ramna*, *Stankena* and *Zablitsena*. More precisely, 1 video from each dance performed by professional dancing groups indoors was used for training and the other one was used for testing. The training videos of the 5 dances were temporally segmented in a manual way into overlapping clips of duration 80 to 100 frames each resulting to 78, 113, 110, 95 and 101 clips respectively. The same procedure was used for the testing videos resulting to 102, 107, 110, 106 and 91 clips respectively. Thus the training and test set consisted of 496 and 516 short sequences respectively. The overall correct classification rate for the three types of features (STIP, ISA and TRAJ) and for various numbers of clusters of the K-means algorithm are shown in Table 1.

Table 1. Performance of STIP, TRAJ and ISA features on the folk dances dataset

Number of clusters	STIP	TRAJ	ISA
10	**49.61**%	32%	54.84%
100	37.60%	35.85%	**78.68**%
1000	38.18%	**44.60**%	72.86%

As can be seen in this table the features learned via deep learning techniques (ISA) clearly outperform the other two features. Considering the difficulties of traditional dance recognition the classification rate achieved with the use of ISA features (78.68%) is very satisfactory.

The histograms produced by the three features were fused to check if the fusion brings performance gains. As already mentioned, the histograms were fused in two ways: either by addition of the χ^2 kernels computed for the initial histograms or by the concatenation of the initial histograms before the kernel computation. The highest correct classification rates for all combinations of the three features are shown in Table 2.

As can be seen in Table 2 the fusion by adding the SVM kernels achieved better classification rate than the fusion by concatenation. The overall best classification rate (79.96%) is achieved by the combination of all three features through kernel addition. However, this classification rate is only slightly better than that achieved by using only ISA features (78.68%). It can also be seen that the classification rate is high whenever ISA features are involved in the fusion. These observations indicate that ISA features are the most suitable ones for the recognition of traditional Greek dances and that their individual classification rate is already high enough to be significantly improved by fusing them with the other

Table 2. Best classification rates for all features combinations

Fused Features	Classification Rate	Fusion Method
STIP/TRAJ	52.71%	adding kernels
ISA/STIP	78.69%	
ISA/TRAJ	78.68%	
ISA/STIP/TRAJ	**79.26**%	
STIP/TRAJ	52.51%	concatenation
ISA/STIP	78.69%	
ISA/TRAJ	72.29%	
ISA/STIP/TRAJ	73.04%	

types of features. When features with lower performance are fused (STIP and TRAJ features) the performance increases by a larger margin (52.71% compared to 49.61% and 44.6% for STIP and TRAJ features respectively). However the performance remains low. This fact can also be seen in Table 3, where STIP and TRAJ feature histograms obtained with the best parameters are fused with ISA feature histograms with low classification rate. The fused classification rate increases to 61.24%. Thus, it is fair enough to assume that the fusion of features is desirable only when the initial features have similar performance.

Table 3. Comparison of the fused classification rate with the initial features when their classification rate is low

	ISA	**STIP**	**TRAJ**	Fused (kernel addition)
Class. Rate/Nr. of clusters	54.84%/10	49.61%/10	44.60%/1000	**61.24**%

The confusion matrix of the best classification rate can be seen in Table 4.

Table 4. Confusion matrix of fused features via adding kernels (ISA, STIP, TRAJ)

	Lotzia	**Capetan Loukas**	**Ramna**	**Stankena**	**Zablitsena**
Lotzia	**95.10**	4.90	0	0	0
Capetan Loukas	7.48	**91.59**	0	0.93	0
Ramna	10	19.10	**70.91**	0	0
Stankena	0	3.77	0	**57.55**	38.68
Zablitsena	2.20	15.38	0	0	**82.42**

As can be seen in this table *Capetan Loukas* and *Lotzia* are recognized with high recognition rates while *Stankena* achieves the worst recognition rate.

It should be noted that, STIP and TRAJ features have been tested on various databases in generic activity recognition with very good results [13], [11]. The low classification rates of these features in the presented experimental setup proves that traditional dance recognition bears significant difficulties making recognition, even between relatively few classes (5), difficult.

5 Conclusions

In this paper we deal with the recognition of Greek traditional dances. Three state of the art methods for feature extraction are used, fused and compared within a bag of words approach. The method is applied on five traditional dances from the Western Macedonia region. The results on these challenging videos are promising but also prove that traditional dance recognition is a very difficult task. STIP and TRAJ features fail to achieve high classification rates. On the other hand the high classification rate of ISA features indicates that features learned through neural networks can be successfully used to recognize videos from Greek traditional dances. Feature fusion provided no significant improvement to the already high performance of ISA features. In the future, we plan to test the presented approach in more rich datasets.

Acknowledgements. 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.

References

1. Poppe, R.: A survey on vision-based human action recognition. Image Vision Comput. 28, 976–990 (2010)
2. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine Recognition of Human Activities: A Survey. IEEE T. Circ. Syst. Vid. 18, 1473–1488 (2008)
3. Xiaofei, J., Honghai, L.: Advances in View-Invariant Human Motion Analysis: A Review. IEEE T. Syst. Man. Cy. C 40, 13–24 (2010)
4. Samanta, S., Purkait, P., Chanda, B.: Indian Classical Dance classification by learning dance pose bases. In: 2012 IEEE Workshop on the Applications of Computer Vision, pp. 265–270. IEEE Press, Washington, DC (2012)
5. Raptis, M., Kirovski, D., Hoppe, H.: Real-time classification of dance gestures from skeleton animation. In: 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 147–156. ACM, New York (2011)
6. Deng, L., Leung, H., Gu, N., Yang, Y.: Recognizing Dance Motions with Segmental SVD. In: 20th International Conference on Pattern Recognition (ICPR), pp. 1537–1540. IEEE Press, Istanbul (2010)
7. Bo, P., Gang, Q.: Binocular dance pose recognition and body orientation estimation via multilinear analysis. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8. IEEE Press, Anchorage (2008)
8. Feng, G., Gang, Q.: Dance posture recognition using wide-baseline orthogonal stereo cameras. In: 7th International Conference on Automatic Face and Gesture Recognition, pp. 481–486. IEEE Press, Southampton (2006)
9. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3361–3368. IEEE Press, Colorado (2011)

10. Laptev, I., Lindeberg, T.: Space-Time Interest Points. In: International Conference on Computer Vision (ICCV), Nice, pp. 432–439 (2003)
11. Wang, H., Kläser, A., Schmid, C., Liu, C.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vision 103, 60–79 (2013)
12. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference, London, p. 127 (2009)
13. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE Press, Anchorage (2008)
14. Shi, J., Tomasi, C.: Good Features to Track. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 1994, pp. 593–600. IEEE Press, Seattle (1994)
15. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part II. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)