Collaborative Activities Understanding from 3D Data

Fabrizio Natola, Valsamis Ntouskos and Fiora Pirri
ALCOR Lab, Department of Computer, Control and Management Engineering 'Antonio Ruberti',
Sapienza University of Rome, Rome, Italy

1 STAGE OF THE RESEARCH

The aim of this work is the recognition of activities performed by collaborating people, starting from a 3D data sequence (specifically, a MOCAP sequence), independently from the point of view from which the sequence is taken and from the physical aspects of the subjects. Many progresses have been made in this field and, in particular, we start from results, for the recognition of actions performed by a single person, obtained in the past years, by Medioni, Gong and other authors, (Gong et al., 2014; Gong and Medioni, 2011; Gong et al., 2012), trying to go a step forward. Following their works, actions are represented by structured multivariate time series in the joint-trajectories space. Two main methods are considered for solving this problem. First, given a sequence consisting of an arbitrary number of actions, *Kernelized Temporal Cut* is applied to find the time instants in which action transitions occur. Then, the *spatio-temporal manifold model*, a framework designed by (Gong and Medioni, 2011), is used for representing the time series data in a one-dimensional space and the spatial-temporal alignment algorithm is introduced in order to find matches between action segments. The resulting procedure, named *Dynamic Manifold Warping*, allows us to classify actions, just by comparing the current action segment with few labelled sequences taken from a given database. The works cited above omit several issues and, without a clear parametrization of them, no implementation is possible. We have sorted out the most suitable parametrization for our implemented algorithm obtaining a good performance in the experiments.

2 OUTLINE OF OBJECTIVES

The main objectives of our work consist in introducing a model that learns the parameters of a distance function on the manifold of the activity sequences. This function would allow the recognition model to generalize from experience, improving the learning. On these bases, we will be able to pass from the recognition of activities performed by a single person to the recognition of activities acted by collaborating people. Most of the research so far has concentrated on single activities ((Gong et al., 2014; Ning et al., 2008; Li et al., 2010), just for citing few of them). Moreover, some works have focused on problems such as the alignment of sequences in time and/or in space, without focusing on learning a function that allows to generalize the model. In particular, the work developed by Gong et al., (Gong et al., 2014; Gong and Medioni, 2011; Gong et al., 2012), regarding the recognition part, is an instance-based approach, since nearest neighbors is considered. In fact, once the Dynamic Manifold Warping algorithm is applied, we have a distance measure between the current testing sequence and all the labelled sequences maintained in a dataset. The label of the nearest sequence (according to Dynamic Manifold Warping) is then assigned to the testing sequence. Because of this aspect and the design of the spatial-temporal alignment algorithm, only few sequences are needed as labelled (for each type of action) to be compared with the testing sequences. On one hand this property can be accounted as positive, since we do not need a large amount of training sequences. On the other hand, however, an approach of this kind has some drawbacks. In fact, the classification in time depends on the number of training sequences and we may obtain wrong answers at query time because of irrelevant attributes. Moreover, the approach does not provide an explicit model for each action class, and therefore it cannot generalize based on the training sequences.

3 RESEARCH PROBLEM

The research problem is twofold, since to comply with our objectives, some basic formal results are needed. Indeed, to model the sought-after recognition of complex actions, such as the collaboration of two people, we first have to provide a representation of the sequences that is flexible and easy to access by the

Natola F., Ntouksos V. and Pirri F..
Collaborative Activities Understanding from 3D Data.

recognition process. We search for a space in which we can map the frame points (i.e. points corresponding to the poses of the human skeleton over time) and in which we do not loose information regarding action sequences performed by several people.

In section 4, we describe related works regarding human activity recognition. In section 5, we explain our current research and possible solutions to overcome the problems mentioned above.

4 STATE OF THE ART

The problem of human activity recognition has been treated in many works from different points of view such as considering, for example, stochastic or non-parametric models.

First, we mention some works which concern the activity recognition starting from video sequences. In (Junejo et al., 2011), the authors consider the action recognition from videos acquired from multiple cameras, studying the similarities of activities over time (building a so-called *Self-Similarity* matrix), using some low-level features (e.g. point-trajectories or HOG descriptors).

In (Weinland et al., 2010), *3D HOG* descriptors are extracted from the training data for obtaining recognition which is robust to occlusions and independent from the point of view from which the scene is observed.

Concerning MOCAP sequences, a graph model called *Action Net* is considered in (Lv and Nevatia, 2007). Each node in the graph represents a 2D rendered figure (called *key pose*) of a MOCAP pose taken from a single view. An edge in the graph, between two nodes, indicates that the two key poses are correlated, in the sense that it represents a transition within a single action class or a transition between two different actions. The Shape Context, which acts as a descriptor of the human silhouette, is considered and stored in each node of the graph. The rendering process is however a drawback if we consider the memory and time required for this operation.

An other important work is (Lv and Nevatia, 2006), in which the authors decompose the 3D joint position space into a set of feature spaces. Each of these subspaces represents the motion of a single part of the body, or the combination of multiple ones. The dynamic information is learned by means of *Hidden Markov Models* and, in order to improve the accuracy of the recognition phase, *AdaBoost* is applied.

There are several works that are strictly related to the topic of study discussed here. In fact, some approaches try to find a low dimensional manifold on which the human motions lie, which they then use in order to solve the problem of recognition. By finding a mapping function from the high dimensional space of the motion to the manifold, it is possible to reduce the dimensionality of the problem.

In (Liao and Medioni, 2008), the authors make use of Tensor Voting for tracking faces in 3D and deducing the facial expressions. We mention this work, even if it is not related to activity recognition, because it shares many similarities with the approach proposed in (Gong et al., 2014) for constructing the *spatio-temporal manifold*. In this latter work, Tensor Voting (Mordohai and Medioni, 2010) is applied for learning the one dimensional path along the spatial manifold on which the frame points lie, consisting the spatio-temporal manifold. By inferring a latent variable which parameterizes this path, it is then possible to apply the temporal and spatial alignments (by means of Dynamic Manifold Warping) for obtaining the distance measure between pairs of sequences and understanding how similar they are.

Another way of modeling a human motion is described in (Zhang and Fan, 2011). Here, the authors consider an action to be modeled using a toroidal two-dimensional manifold in which the horizontal and vertical circles represent two different variables: gait and pose, respectively. This manifold can be identified by considering a *joint gait-pose manifold* (JGPM), which is based on *Gaussian Process Latent Variable Model* (GP-LVM), (Lawrence, 2004).

Finally, in (Ntouskos et al., 2013), sequence alignment kernels (Noma, 2002; Cuturi et al., 2007) are combined with *Back-Constrained GP-LVM* (Lawrence and Quinonero-Candela, 2006), for achieving action recognition. Also in this case, a dimensionality reduction is performed for mapping MOCAP data into a lower-dimensional manifold.

5 METHODOLOGY

In this section, we concentrate on discussing the problems that we have noticed, proposing different contributions to improve the work of Gong et al. In the first part, we explain the problem of constructing a learning function from which we can generalize the model of (Gong et al., 2014), and in the second part we give possible research directions in order to achieve recognition of activities made by two people that work together.

5.1 Generalization of the Model

Following the idea of Gong et al. in (Gong et al., 2014), we consider a MOCAP sequence as a multivariate structured time series. In this sense, the sequence can be encoded in a matrix. Namely, the matrix is constructed in the following way:

$$M = [x_1 \ \ x_2 \ \ \cdots \ \ x_L],$$

where L is the length of the sequence in terms of number of frames and each x_i, for $i = 1, \cdots, L$, represents the joint positions at time i. Namely:

$$x_i = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{bmatrix},$$

where J is the number of joints considered in the MOCAP skeleton, and each p_j, for $j = 1, \cdots, J$, is the position of the j-th joint. Each p_j is in turn composed by its x, y, z positions:

$$p_j = \begin{bmatrix} p_j^x \\ p_j^y \\ p_j^z \end{bmatrix}.$$

Therefore, each frame is considered as a point in a high dimensional space. In detail, the space has dimension D, where $D = 3 \times J$. However, we can make some assumptions that are useful for reducing the dimensionality of the problem. First of all, skeleton motion is constrained in time. Secondly, if we move a limb, it is very likely that other limbs linked to the former will move also. Because of these constraints, we can safely say that these points lie on a manifold having dimension $S < D$.

In order to obtain a learning function, we cannot directly apply learning methods such as the *Support Vector Machines* (SVM) or *discriminant analysis*, because the points lie on a non Euclidean space. We need to find a suitable transformation that allows us to apply these methods in our case. (Vemulapalli et al., 2013), for example, use kernel functions for mapping the points on a Riemannian manifold into a *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}$. This latter is a Hilbert space endowed with the reproducing property:

$$\langle f(\cdot), k(\cdot, x) \rangle = f(x)$$

for each function $f \in \mathcal{H}$ and every $x \in \mathcal{X}$, where $\mathcal{X}$ is the domain of the variable x and $\langle \cdot, \cdot \rangle$ is the inner product operator. Consequently, we can derive:

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}},$$

where $\Phi(x) = k(\cdot, x)$ is the feature mapping function from $\mathcal{X}$ to $\mathcal{H}$. This transformation allows to apply one of the classical learning algorithms, such as SVM, for constructing a learning function needed for the classification process. An important question pointed out by Vemulapalli et al. in (Vemulapalli et al., 2013) is how to select a good kernel for mapping the points on the manifold into the RKHS. The answer comes from (Rakotomamonjy et al., 2008), in which the authors construct a kernel as a linear combination of simpler base kernels:

$$K = \sum_{i=1}^{M} \mu_i K_i,$$

with $\mu \in \mathbb{R}^M$ being a vector of positive weights that must be found, and K_i being the base kernels. Therefore the problem is to jointly optimize the classifier problem and the kernel problem:

$$\min_{W,K} \lambda \mathcal{L}_M(K) + \mathcal{L}_C(W,K),$$

where $\mathcal{L}_M(K)$ is the manifold structure cost, while $\mathcal{L}_C(W,K)$ is the classifier cost. These are functions of the arguments of the minimization W and K, which are in turn the parameters for the classifier and the kernel function, respectively.

5.2 Recognition of Activities Made by Two Collaborating People

As mentioned before, many progresses have been made in human action recognition based on labelled MOCAP sequences, but so far temporal analysis has mainly focused on sequences corresponding to a single subject, even of several subsequent activities. For a single subject sequence the temporal segmentation problem is about identifying the correct subdivision of all the activities, up to single actions. Significant contributions in this direction are, for example (Ali and Shah, 2010; Gong et al., 2014; Gong et al., 2012). With a single MOCAP sequence, temporal segmentation takes care of the chain of actions to feed the recognition process, and both temporal manifold and spatio-temporal alignments do not consider composite motion of more than a person. Moreover, as temporal motion is highly varying amid different people, the problem cannot be simply lifted to two different motions acting in parallel. In fact this problem requires the existence not only of two temporal sequences, but also constraints between them which model the interaction (Li et al., 2013; Ryoo and Aggarwal, 2011).

After having generalized the model and having constructed a suitable learning function, we go a step forward and consider the recognition of the interleaving process between two action sequences. We assume that labelled MOCAP data are provided (see the image sequence in Figure 1) and the objective is to predict the interleaving steps of the two processes. The main contribution is the spatio-temporal alignment of the two motion sequences so as to identify when and how the two sequences intertwine due to an action requiring the collaboration of two subjects.

In particular we are interested in the collaboration in a working environment between two operators executing a task that requires to handle tools, pass them and, possibly, handling together some item, in which case the spatio-temporal alignment is crucial. We consider motion sequences of industrial related activities performed by human operators, as for example maintenance and repair operations. We focus on learning interaction patterns involved in collaboration type of actions, in order to be able to identify the time instances where physical interaction occurs. These time instances serve as nodal points which determine the evolution of the whole interactive activity. Identifying these nodal points is a crucial issue. In fact these points may vanish by using solely Latent Variable Models (Tenenbaum et al., 2000; Roweis and Saul, 2000; Wang et al., 2008; Mordohai and Medioni, 2010) during the mapping from the ambient space to the intrinsic, lower-dimensional space. In order to avoid this, we shall consider three sequences: one for each person involved and one for the interleaved action. By learning the nodal points involved, we can learn a model of the interleaved action. In the case of simultaneous handling of items, the evolution of the action between the nodal points also plays an important role. In particular, spatio-temporal alignment with respect to prototypical evolutions of the action is necessary in order to be able to maintain continuous interaction and evaluate its dynamics. The model of the interleaved action can be later used for enabling analogous human-robot interactions. We can think that the role of one of the two people involved can be substituted by a robot. In this latter case, the robot has to understand the motion of the person it is assisting, and it has to decide which action it has to perform.

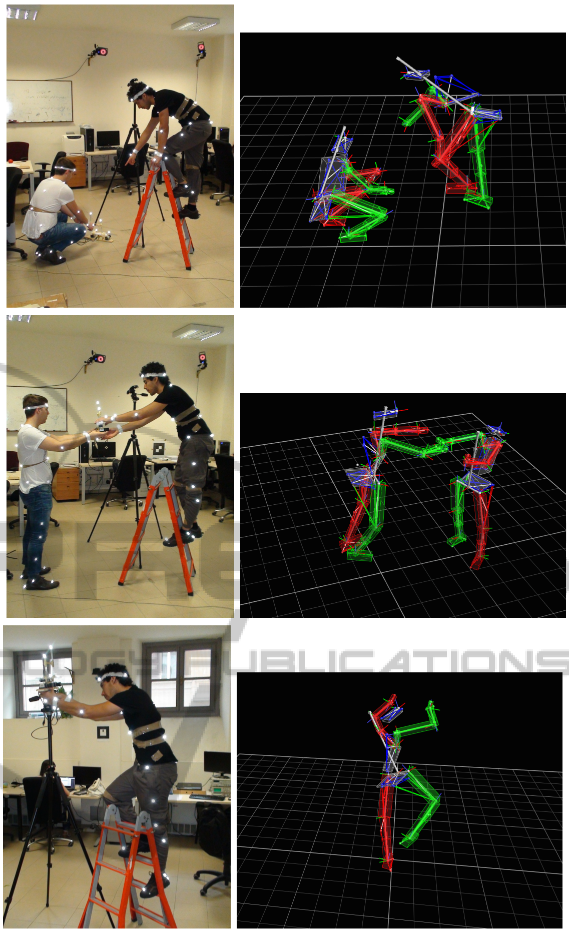

Figure 1: A collaboration action sequence illustrating two subjects executing an interleaving task. *Left col.*: Instances of the action; *Right col.*: Corresponding MOCAP poses. We consider as input 3D points, that in our case are obtained by a Vicon system, but they can be also computed from a depth video.

6 EXPECTED OUTCOME

By implementing our version of the framework proposed by Gong et al. in (Gong et al., 2014), we have achieved high accuracy in recognition by considering MOCAP skeletons composed by 31 joints. In particular, we have taken into account 7 actions (i.e. squat, run, hop, walk, kick, sit down, rotate arms), each of which performed by 5 different subjects, for a total of 35 sequences. Each sequence is compared with all the other ones and the accuracy is computed as the number of right recognized sequences over the total number of sequences. The resulting accuracy is equal to 86% suggesting that the algorithm recognizes a significant number of sequences.

We noticed that even in the presence of some joints that may introduce noise (such as near the hands and the feet), the accuracy is very satisfactory. We expect to generalize the model and to construct a learning function capable of representing distances in a space where actions performed by different people can be represented.

REFERENCES

Ali, S. and Shah, M. (2010). Human action recognition in videos using kinematic features and multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):288–303.

Cuturi, M., Vert, J., Birkenes, O., and Matsui, T. (2007). A kernel for time series based on global alignments. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–413–II–416.

Gong, D. and Medioni, G. (2011). Dynamic manifold warping for view invariant action recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 571–578.

Gong, D., Medioni, G., and Zhao, X. (2014). Structured time series analysis for human action segmentation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1414–1427.

Gong, D., Medioni, G., Zhu, S., and Zhao, X. (2012). Kernelized temporal cut for online temporal segmentation and recognition. In *Computer Vision–ECCV 2012*, pages 229–243. Springer.

Junejo, I., Dexter, E., Laptev, I., and Perez, P. (2011). View-independent action recognition from temporal self-similarities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):172–185.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16:329–336.

Lawrence, N. D. and Quinonero-Candela, J. (2006). Local distance preservation in the gp-lvm through back constraints. In *Proceedings of the 23rd international conference on Machine learning*, pages 513–520. ACM.

Li, R., Chellappa, R., and Zhou, S. K. (2013). Recognizing interactive group activities using temporal interaction matrices and their riemannian statistics. *International journal of computer vision*, 101(2):305–328.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14.

Liao, W.-K. and Medioni, G. (2008). 3d face tracking and expression inference from a 2d sequence using manifold learning. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Computer Vision–ECCV 2006*, pages 359–372. Springer.

Lv, F. and Nevatia, R. (2007). Single view human action recognition using key pose matching and viterbi path searching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8.

Mordohai, P. and Medioni, G. (2010). Dimensionality estimation, manifold learning and function approximation using tensor voting. *The Journal of Machine Learning Research*, 11:411–450.

Ning, H., Xu, W., Gong, Y., and Huang, T. (2008). Latent pose estimator for continuous action recognition. In *Computer Vision–ECCV 2008*, pages 419–433. Springer.

Noma, H. S. K.-i. (2002). Dynamic time-alignment kernel in support vector machine. *Advances in neural information processing systems*, 14:921.

Ntouskos, V., Papadakis, P., Pirri, F., et al. (2013). Discriminative sequence back-constrained gp-lvm for mocap based action recognition. In *International Conference on Pattern Recognition Applications and Methods*.

Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y., et al. (2008). Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Ryoo, M. and Aggarwal, J. (2011). Stochastic representation and recognition of high-level group activities. *International journal of computer Vision*, 93(2):183–200.

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

Vemulapalli, R., Pillai, J., and Chellappa, R. (2013). Kernel learning for extrinsic classification of manifold features. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1782–1789.

Wang, J., Fleet, D., and Hertzmann, A. (2008). Gaussian process dynamical models for human motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):283–298.

Weinland, D., Özuysal, M., and Fua, P. (2010). Making action recognition robust to occlusions and viewpoint changes. In *Computer Vision–ECCV 2010*, pages 635–648. Springer.

Zhang, X. and Fan, G. (2011). Joint gait-pose manifold for video-based human motion estimation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 47–54.