Proposal for a Deep Learning Architecture for Activity Recognition

Ruben Glatt, José C. Freire, Jr., and Daniel J. B. S. Sampaio

***Abstract*— Activity recognition from computer vision plays an important role in research towards applications like human computer interfaces, intelligent environments, surveillance or medical systems. In this paper, we propose a gesture recognition system based on a deep learning architecture and show how it performs when trained with changing multimodal input data on an Italian sign language dataset. The results show the importance of choosing the right data representation for activity recognition tasks.**

***Index Term*— Activity recognition, Computer vision, Deep learning, Multimodal learning**

I. Introduction

THE focus of this work is on multiple instance, user independent learning of gestures from multi-modal data, which means learning to recognize gestures from several instances for a number of categories performed by different actors. It proposes a learning architecture for gesture recognition using deep learning principles on multimodal data inputs.

Human activity recognition is playing an active role in today's research efforts on still images and video sequences. The research area covers a wide field of complex and challenging problem settings. The three most important stages are segmentation, feature selection and behaviour or action recognition. Segmentation is performed on each single image or frame of the sequence to identify and highlight the main areas of interest. The next step includes the determination and selection of suitable characteristics in the image like colours, shapes, edges or geometric forms, which represent the features of the image, in the case of videos this can be summarized as space-time shapes for example. The recognition step uses those feature representations of the original data in a learning model to categorize the captured scene. The goal of the research is to automatically analyse scenes and provide context and meaning to the recognized activities.

This work is inspired by the "ChaLearn 2014 - Looking at people" challenge, which focuses on multiple instance, user in- dependent learning of gestures from multi-modal data. The authors want to thank the organizers of the challenge for generating the dataset and providing it to download.

Ruben Glatt is with the Mechanical Engineering Department, Universidade Estadual Paulista - Campus Guaratinguetá, Brazil (e-mail: ruben.glatt@feg.unesp.br)

José C. Freire, Jr., and Daniel J. B. S. Sampaio are with the Electrical Engineering Department, Universidade Estadual Paulista - Campus Guaratinguetá, Brazil (e-mail: : jcfreire@feg.unesp.br, dsampaio@ feg.unesp.br).

II. Related Work

Gesture or activity recognition is a prevailing topic in artificial intelligence and machine learning research. In recent years devices for capturing 3D data have been made available for a wide range of people through sensors like Microsofts Kinect or the Primesense depth sensor. Beside the many available applications, these low cost sensors provide a great basis for research in computer vision and activity recognition.

Before the commonness of depth sensors, research was mostly limited to analysing RGB data. Comprehensive summaries of various approaches with RGB data to human motion capture, analysis, modelling, initialization, tracking, pose estimation and action recognition are given by Aggarwal and Park (2004) and Moeslund et al (2006).

A popular approach in human activity recognition is to find the human skeleton with central joints or select body parts and analyze the positions towards each other as discussed by Zhuang et al. (1999), Ramanan and Forsyth (2003) and Felzenszwalb and Huttenlocher (2005).

An important factor in recognizing activities through computer vision is the definition of an activity and the meaning of it, which is further described by Krüger et al. (2007).

Denman et al. (2007) propose an optical flow technique that is based upon an adaptive background segmentation technique, which only determines optical flow in regions of motion. Gorelick et al. (2007) regard human action in video sequences as silhouettes of a moving torso and protruding limbs undergoing articulated motion and generate three-dimensional shapes induced by the silhouettes in the space-time volume. Niebles et al. (2008) propose an unsupervised learning method for human action categorization where they extend the Bag-of-Words (BoW) approach of still images to introduce Spatio-Temporal-Visual-Words (STVW).

A survey by Aggarwal and Ryoo (2011) provides a detailed overview of various state-of-the-art research papers on human activity recognition. It discusses the methodologies developed for simple human actions and for high-level activities and compares the advantages and limitations of each approach. The new generation of visual sensors generated the possibility to explore a wider area of machine vision and integrate new approaches. Schwarz et al. (2011) build upon a graph-based representation of depth data to measure geodesic distances between body parts instead of relying on appearance-based features for interest point detection to avoid illumination and

pose changes influence. Shotton et al. (2011) present a method for body joint detection in single depth images to estimate body parts invariant to pose, body shape, clothing, etc. They treat the segmentation into body parts as a per-pixel classification task and train a deep randomized decision forest classifier, which avoids over-fitting by the huge amount of training data they feed into it. The resulting 3D joint proposals are then computed using mean shift on the spatial modes of the inferred per-pixel distributions. Raptis and Sigal (2013) use video sequences with temporally local discriminative keyframes to capture gestures while Oreifej and Liu (2013) present a new descriptor for activity recognition from videos with depth information using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates.

The review on video-based Human Activity Recognition by Ke et al. (2013) discusses general applications and core technologies and provides a state-of-the-art review of the field.

III. Dataset

The dataset for the recognition algorithm is taken from Track 3 of the "ChaLearn 2014 - Looking at people" challenge. It offers a total of 13.858 gestures, recorded with a Microsoft Kinect. The gestures are split in three sets, which consists of a training set of 7.754 gestures, a validation set of 3.362 gestures and an evaluation set of 2.746 gestures. Each gesture is provided in several formats as a sequence of image frames and taken from a vocabulary of 20 Italian sign language gestures presented by different actors. The 20 gestures are shown in Fig.1 with the underlying meaning of each gesture in Table 1.

The dataset provides each gesture capture from a fixed camera position with near frontal view acquisition of a single actor in the picture. Each sequence features only one actor and each gesture has several instances. The gestures are being performed in a continuous manner without a resting pose and the most active limbs are the arms and hands. The shots of the different actors vary in background, clothing, skin colour and lighting with occasionally occluded body parts.

The multimodal data streams for each gesture consist of the following channels:

- Ground truth: CSV file with the ground truth. Each line corresponds to the sequence for one gesture. Information provided is the gesture truth, the initial frame and the last frame of the gesture.
- RGB: 8 bit, VGA resolution with a Bayer colour filter, 20 fps, mp4
- RGB-D: 11 bit, VGA resolution, 20 fps, mp4
- User: 8 bit, VGA resolution of silhouette data of the actor, 20 fps, mp4
- Skeleton: CSV file with skeleton information for each frame of the videos as illustrated in Fig.2. Each line corresponds to one frame. Skeletons are encoded as a sequence of joints, providing 9 values per joint:
  - world coordinates: global position of joint in mm (W_x, W_y, W_z)
  - rotation values: Quaternion to encode the axis–angle representation in four numbers, and to apply the corresponding rotation to a position vector representing a point relative to the 'HipCenter' joint (R_x, R_y, R_z, R_w)
  - pixel coordinates: position of the joint mapped to the 2D frame in pixel (P_x, P_y)

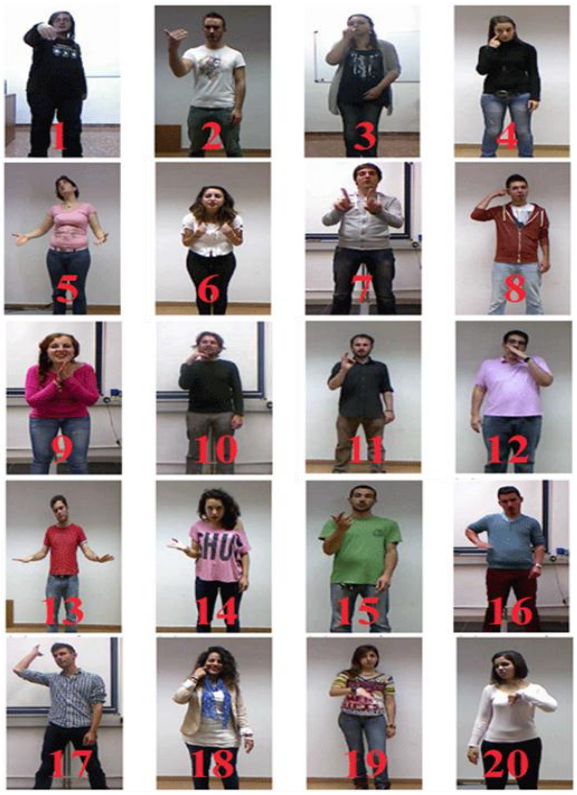

Fig. 1. Exemplary visualization of the 20 different gestures of the dataset

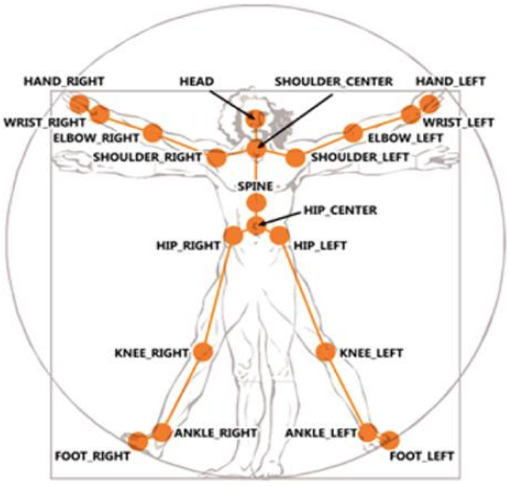

Fig. 2. All available joint indicators

IV. Proposed Learning Architecture

A. Feature extraction

For this work, we used two kinds of feature extraction and further detected the dominant hand of the actor by tracking the traveling distance of the hand joint in the 3D space. Using combinations of these techniques gave us a total of six different approaches to feed into the learning network.

The first method we used was applied to the Skeleton data of the recorded gestures. We identified the most relevant joints ('ShoulderLeft', 'ElbowLeft', 'WristLeft' ,'HandLeft', 'ShoulderRight', 'ElbowRight', 'WristRight', 'HandRight', 'ShoulderCenter', 'Head' ,'Spine', 'HipLeft', 'HipCenter', 'HipRight') and tracked their movement throughout the sequence. Additionally the distances to a central point, the 'Spine', were calculated. All the values were normalized in each frame to decrease the influence of different sizes of the actors or distance to the camera of the actor. At the end of each sequence it was resized to 40 frames to decrease the influence of different moving speeds of the actors. An example is shown in Fig.3 were a short sequence was stretched.

The second method used was applied to the RGB data of the recorded gestures. The sequences are summarized in Motion History Images where the brightness of the pixel decreases for past frames. This way all the frames of a sequence are represented in a final image and the movement during the sequence is captured. The Motion History Image $H_\tau(x,y,t)$ is defined in (1).

$$H_\tau(x,y,t) = \begin{cases} \tau, & if\ I(x,y,t)=1 \\ \max(H_\tau(x,y,t-1)-1), & otherwise \end{cases} \quad (1)$$

where I(x,y,t) is a binary image generated through frame subtraction to highlight movement in the image. The final Motion History Image was then reduced in size to reduce the input data for the learning network. An example of the resulting image is given in Fig. 4.

In the experiments were the dominant hand detection was used, the resulting images or distances from the Skeleton data were mirrored so it appeared to present only hand movement on the right hand side. This technique was used to generate more examples for the same gesture without the diversion of different handedness.

B. Deep learning network

An implementation of a Deep Learning system are Deep Belief Networks. At its introduction by Hinton et al. (2006), Deep Belief Networks were a milestone for Machine Learning applications and rapidly followed by successes on many standard problems. Solutions based on the technique improved existing solutions for example in feature extraction on various datasets by Ranzato et al. (2007), on large unlabelled data sets by Salakhutdinov and Hinton (2008), in long-range vision for autonomous off-road driving by Hadsell et al. (2008) and more.

TABLE I
Vocabulary of Used Signs in the Dataset

Sign	Italian	English
1	vattene	begone
2	vieni qui	come here
3	perfetto	perfect
4	e un furbo	clever one
5	che due palle	how boring
6	che vuoi	what do you want?
7	vanno d'accordo	get along
8	sei pazzo	you are crazy
9	cos'hai combinato	what have you done?
10	non me ne frega mente	I don't care
11	ok	ok
12	cosa ti farei	what would you do?
13	basta	it's over
14	le vuoi prendere	you want to take it?
15	Non ce ne piu	there is no more
16	ho fame	I am hungry
17	tanto tempo fa	a long time ago
18	buonissimo	delicious
19	Si sono messi d'accordo	do we agree?
20	Sono stufo	I'm tired

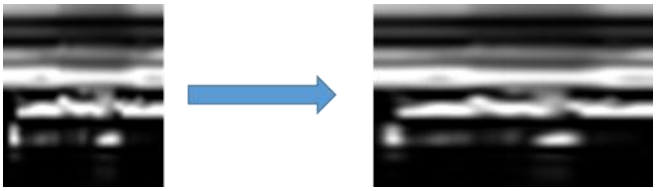

Fig. 3. Visualisation of movement and distance indication of selected joints

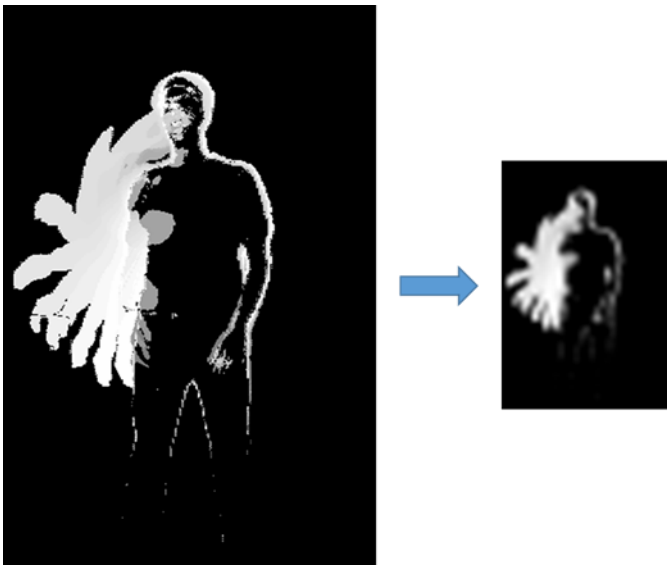

Fig. 4. Example of a Motion History Image in original size and reduced

A Deep Belief Net can be viewed as a composition of simple learning modules each of which is a restricted type of Boltzmann Machine that contains a layer of visible units that represent the data and a layer of hidden units that learn to represent features that capture higher-order correlations in the

data. The two layers are connected by a matrix of symmetrically weighted connections and there are no connections within a layer.

The Restricted Boltzmann Machines of a Deep Belief Network are trained each layer at a time. The structure of a Deep Belief Network with three hidden layers is illustrated in Fig. 5. In the first phase (Fig 5. a) a Restricted Boltzmann Machine is trained using the input data and the first hidden layer, in the second step (Fig. 5. b) the hidden layer of the first step is taken as input of the second Restricted Boltzmann Machine with a second hidden layer on top. In the last step (Fig. 5. c) the hidden layer of a third Restricted Boltzmann Machine is concatenated with the target classification to initialize the weights of the highest hidden layer for the training process as described in Larochelle, et al. (2007).

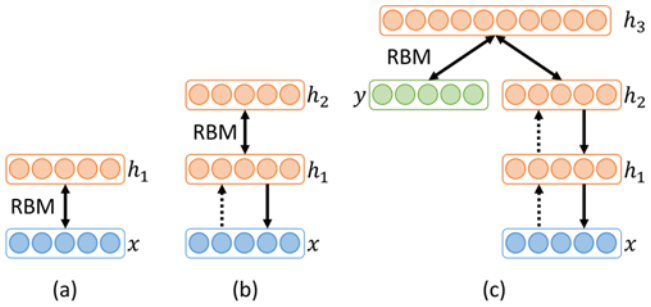

Fig. 5. Structure of a Deep Belief Network

V. Experiment Description

The results provided in the Table 2 were gained from using only a small subset of the provided dataset to get tangible results without running into computational restraints. The training set was composed of 1256 randomly selected sequences, the validation set of 293 sequences and the final test set of 300 sequences of the original data set. The values represent the error rate on the test set compared to the available ground truth data.

VI. Conclusion and Future Work

The results achieved with only a subset of the original dataset are not yet promising and still fail to compare favorably to state-of-the-art machine learning approaches.

An interesting aspect of the results is that obviously the determination of the dominant hand does not improve the recognition rate.

While Deep Belief Networks already achieve very good results on datasets like MNIST, the setup proposed here should be further developed and enhanced. Future work needs to better adjust the proposed architecture for activity recognition from a sequence of images. In the eyes of the authors a key element in achieving better results is finding a better representation of the original data to feed into the network for example with Dynamic Time Warping.

Another way to improve the results could be to use all the available data from the dataset to train and fine-tune the network with more examples.

A stronger infrastructure would allow building a bigger network with more nodes, but the data structure would become very complex and computational limits would become an issue with increasing network size.

TABLE II
Results for Subset of Provided Data

Method	Without dominant hand detection	With dominant hand detection
Skeleton	42,33 %	44,33 %
MHI	51,67 %	57,33 %
Skeleton + MHI	**38,00 %**	41,67 %

References

[1] J. K. Aggarwal and S. Park, "Human Motion: Modeling and Recognition of Actions and Interactions," in *Proc. 2nd International Symposium on 3DPVT, 2004*, pp. 640–647.

[2] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, art. 16, 2011.

[3] S. Denman, V. Chandran and S. Sridharan, "An adaptive optical flow technique for person tracking systems," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1232-1239, 2007.

[4] P. F. Felszenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, pp. 55-79, 2005.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as Space-Time Shapes,". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, nr. 12, pp. 2247-2253, 2007.

[6] R. Hadsell, A. Erkan, P. Sermanet, M. Scoffier, U. Muller and Y. LeCun, Y., "Deep belief net learning in a long-range vision system for autonomous off- road driving," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 628-633, 2008.

[7] G. E. Hinton, S. Osindero and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, nr. 7, pp. 1527-1554, 2006.

[8] S.-R. Ke. H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang,J.-H. Yoo and K.-H. Choi, "A Review on Video-Based Human Activity Recognition," *Computers*, vol. 2, manuscripts, pp. 88-131, 2013.

[9] V. Krüger, D. Kragic, A. Ude and C. Geib, "The meaning of action: A review on action recognition and mapping," *Advanced Robotics*, vol. 21, p. 1473–1501, 2007.

[10] H. Larochelle, D. Erhan, A. Courville, J. Bergstra and Y. Bengio, "An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation," *24th International Conference on Machine Learning*, pp. 473-480, 2007.

[11] T. B. Moeslund, A. Hilton and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," In *Computer Vision and Image Understanding*, vol. 104, p. 90–126, 2006.

[12] J. C. Niebles, H. Wang and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, vol. 79, nr. 3, pp. 299–318, 2008.

[13] O. Oreifej and Z. Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716-723, 2013.

[14] D. Ramanan and D. Forsyth, "Finding and tracking people from the bottom up," in *Proc. Computer Vision and Pattern Recognition, 2003*, vol. 2, pp. 467–474.

[15] M. A. Ranzato, F. J. Huang, Y.–L. Boureau and Y. LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[16] M. Raptis and L. Sigal, L., "Poselet Key-framing: A Model for Human Activity Recognition," *Conference on Computer Vision and Pattern Recognition*, pp. 2650-2657, 2013.

[17] R. Salakhutdinov and G. E. Hinton, "Using Deep Belief Nets to Learn Covariance Kernels for Gaussian Processes," *Advances in Neural Information Processing Systems*, vol. 20, p. 1249-1256, 2008.
[18] L.A. Schwarz, A. Mkhitaryan, D. Maeus andN. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow," *Image and Vision Computing*, vol. 30, p. 217–226, 2012.
[19] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 3, 2011.
[20] Y. Zhuang, X. Liu and Y. Pan, "Video Motion Capture Using Feature Tracking and Skeleton Reconstruction," *International Conference on Image Processing*, vol. 4, pp. 232 – 236, 1999.