A Novel DWT-based Encoder for Human Pose Estimation

Giorgio De Magistris[1], Matteo Romano[1], Janusz Starczewski[2] and Christian Napoli[1,3]

[1]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, Roma, 00185, Italy
[2]Department of Computational Intelligence, Czstochowa University of Technology, al. Armii Krajowej 36, Częstochowa, 42-200, Poland
[3]Institute for Systems Analysis and Computer Science, Italian National Research Council, Via dei Taurini 19, Roma, 00185, Italy

Abstract

The proposed approach for pose estimation is based on the construction of a Convolutional Neural Network with an encoding-decoding structure and a spatial pyramid based on WASP structure in its bottleneck and a Discrete wavelet transform encoder. These techniques already shown their capabilities to solve the main problems in state of the art related to: different Field of view (FoV) required to analyze the different possible sizes of a specific subject. we want to solve the faulty structure of the modern CNN based Neural Networks in the encoding part using DWT encoder and WASP. This Work also have the objective of demonstrating from a more general point of view which could be the advantages of a Discrete Wavelet Transform (DWT) encoder in any CNN-based approach for Pose Estimation and Object detection in any form, such as for several subjects in the same image or in the internal video due to the almost redundant use of the usual most famous encoding structures for CNN such as ResNet-101, U-Net or VGG16-19. we will do our tests using a U-net Based CNN in order to evaluate the importance of the results of the Discrete Wavelet Transform encoder also in the decoding part through the cropping of theme at the last layers of the network. This is necessary due to the loss of border's pixels during encoding that could be useful for the result's evaluation.

Keywords

Discrete Wavelet Transform, Convolutional Neural Network, WASP, Atrous Convolution

1. Introduction

Pose Estimation task is important for many aspects from Human detection and pose estimation to the navigation system for autonomous car and also different fields as object detection and image segmentation. We can use different types of approaches as Top-down detection human with a bounding box as object detection task and then use the pose estimation algorithm as in [1]. An alternative for bottom-up approach estimate the points from the image and then recreate the human form by the "skeleton" given by the conjunction of these points. The method used in our case will be the Top-down for the estimation of these points. This is because the result that we want to emphasize is the improvement of the performances at encoding level so that include alle the structure of this kind even the ones used for object detection. There are different challenges in the capability of certain CNN to get good results even in complex situations and In a different way the novelty of our study could affect the capabilities of these CNN structures. The actual Convolutional neural network Based method for the key points detection and estimation of the pose of human limbs is actually all developed through encoding-decoding structures which are very similar to each other. More specifically, structures based on a representation of the image in a latent space through the construction of feature maps from a local to a global viewpoint through Spatial Pyramid approach as bottleneck of the CNN. For this type of task, various technologies are currently present in the state of the art for improving the performances of these CNNs, also related to the management of the different fields of view (FoV) of the image necessary for the evaluation of objects with different scales in the representation but what we want to add with this work is the resolution of problems localized in the encoding part of these neural networks. The final product of this kind of structure for pose estimation of a single subject will be at the end a simple set of points to interpolate or a heatmap of the image obtained through the generation in pre-processing phase of a ground truth (GT) image with gaussian noise in the keypoints area for the evaluation of a loss function based on a percentage of Correct Key points respect the GT image and predicted key points. There are actually different challenges in the state of the art not only related the technologies actually used but also problems due to the case of possible occlusion of limbs due to presence of different persons in images and requiring the use of Multi Pose estimation or low-quality of the images as in video where we have very fast objects moving in the image that needs of less smooth corners for the analysis. Other examples where is possible to find that this kind of structure are used beyond the Human pose estimation also for hand pose as in [2, 3, 4] where are analyzed many kinds of wavelet applied to the purpose or even different applications in the state of the art [5, 6] as noticed also

SYSYEM 2022: 8th Scholar's Yearly Symposium of Technology, Engineering and Mathematics, Brunek, July 23, 2022
demagistris@diag.uniroma1.it (G. De Magistris);
janusz.starczewski@pcz.pl (J. Starczewski);
cnapoli@diag.uniroma1.it (C. Napoli)

by different works in image classification as [7, 8] or for object detection approaches as in [1, 9] or other solutions [10, 11]. With this premises our work is focused on the application of an U-net based model with a multi-level decomposition (MLD) of the image as encoder parallel to U-net with concatenation of the gathered information for each layer of the DWT encoder to the encoder of U-net and propagation of this information to the decoder. In the feature extraction at the U-net's bottleneck is instead used a WASP structure to obtain the resulting feature map of the image from different field of view after the encoding part to pass to the decoder. In this way the encoder do not loose the information eliminated by the down sampling operation because reconstructed from the information of wavelets passed forward to the decoder.

2. Related Works

Two main aspect of this works are considered as first respect to the works related to the wavelet applied to different fields as Image segmentation or object detection and then the common structures and modern approaches actually used in Pose Estimation, this work is focused on the analysis of approaches from both this two aspects used in combination with pre-processing common practise related this to the task of Human Pose Estimation. There are many pose estimation systems [12, 13, 14], among them UniPose [15] is a pose estimation method based on the application of a so-called WASP structure as the central part of the bottleneck in the CNN. It is based on the use of layers with different dilation in the application of a rate parameter related to the formula:

$$y[i,j] = \sum_{i=0}^{L_1} \sum_{j=0}^{L_2} I[m - ri, m - rj] * k[i,j] \quad (1)$$

with r = rate of dilation , I = image and k as kernel over the x/y-axis and L_1 or L_2 as pixel's position. In this way is possible to get a higher FoV for the image and connect it to a depth-wise convolution operation to obtain a higher level of abstraction features to consider in the latent representation. UniPose reached a Percentage of Correct Parts (PCP) considerably high for the actual state of the art but not even near to be considered as robust approach respect most complex methods based on 3D models of poses. One of the most important approaches based on Cascaded Pyramid Network and pointing to give a 3D representation of the scene as in [16] however we are not interested to highlight because too much computational expensive and applicable only in specific environments as in the case of multi-camera detection of the scene. A particular explanation of wich problems can be encountered in this kind of task is given by PersonLab [17] where the Atrous approach is also analyzed for multiple subject in the image in which case we not only establish a heatmap for the evaluation of the key points in a Gaussian map but we consider have to also a segmentation of the image and evaluation of long,medium and short range offset for each key point of the subject to establish the relationship between different key points belonging to different subject and the ones in the same segmented part of the image. In these cases is also important to mention as in [18] the WASPv2 version used in combination to HRNet structures without a decoding structure that follow it performing a cascade of Atrous convolutions at increasing rates to gain efficiency. In the next section are explored the past works in DWT application to a variety of state of the arts and explains how our approach is different in chapter 3 and in the conclusion how the results are improved by it. Talking about possible applications of Discrete Wavelet Transform (DWT) related to different state of the art we can consider as fundamental the contribute given by D-Unet a dual encoder used for different purposes as Image segmentation and object detection [19]. DWT-based encoder is an important addition to the state of the art because it demonstrated its superiority in reconstructing information lost in the encoding part of the neural network and also the superior capability with respect to different methods used as steganalysis rich model (SRM). In which case the layer extracts the image noise providing additional evidence for the classification of multiple types of image key points as shoulders, elbows and faces. In [20] has been instead demonstrated the problems related the use of a down-sampling and up-sampling operation and corresponding interpolation operation needed to reconstruct the original image from the global feature map at the start of the decoding part. As last consideration is important to cite also the capabilities demonstrated by spatial pyramid in obtaining of low-resolution feature maps with global features if used with different FoV and dilation parameters the so called Atrous Spatial Pyramid Pooling (ASPP) compared to other simpler architecture to handle different scales, sizes, and aspect ratios of the subjects.

3. Model Architecture

The complete structure of the model is separated in three parts: The encoder and decoder structure typical of U-net with propagation of the information from decoding to encoding layers, the bottleneck where the waterfall Atrous spatial pyramid (WASP) is used to obtain feature map in different Field of view and the parallel encoder used for the multi-level decomposition of the image for each different image's channel with the corresponding concatenation of the low pass representation in the layers of U-net delegated to the forward propagation of

the information to the decoder. The complete architecture is visible in where each block represent a layer of U-Net with: down sampling operation, dropout and normalization of the batch. We selected two datasets for the validation and training the first is COCO containing 40000 images while a more specific and generally challenging dataset is LSP for this kind of task that , used in combination with a small part of COCO to obtain a single dataset of 3600 samples and the remaining for testing and validation. The LSP dataset includes modified data with noise addition, having a good assessment of the network performance for the task of single person pose estimation and even in this case one of the most problems is the occluded limbs.

As mentioned in [21] the Fully Convolutional Networks (FCN) are the most used kind of CNN in this fields and all are structured as encoder-decoder with up-sampling procedure for reconstruction of a resolution and restore of loosed data in encoding part. In this section are considered as assumption that the structure will be similar to U-Net. however is important to remark that these kind of structures already establish state of the art result without thorough consideration of other methods of image feature extraction. As first approach has been considered a solution based on the generation of an heat map but in that case the construction of it is very similar to the binary segmentation of [19] which can lead to problems as the necessity to use heat maps separated for each key point in order to connect theme each other in a correct way even if we have overlap of it. So in this case a simple binary segmentation is not useful for overlap of limbs and articulations and a more computational complexity shows up in working separately for each key point. For these reasons has been chosen to consider a layer for a regression task with a vector representing the scaled coordinates of the points in the image that we have to interpolate with our model. The reason behind the U-net structure as choice, instead most common VGG or ResNet, is due to the possibility to propagate the information produced with wavelet's coefficients everywhere in the layers between encoding and decoding. The network to propagate context information to higher resolution layers, exploiting this capability of Unet the information propagated are in the structure of our Neural Network the information gathered from the wavelet in the encoder layers concatenated layer by layer to the information obtained from the network's encoder. This will lead us to a distance function weighted respect to the subject dimension in the image that we will have to minimize (e.g simple MSE respect the position) and find the best function that interpolate the position of the key-points respect the image's information. We want in this way to be able to build augmented information for an image with very low dimension and use them to infer invisible information for a simple U-net encoder-decoder. WASP is designed with the goal of reduce the number of parameters in order to deal with memory constraints and solve the main issue of Atrous convolutions using different FoV for image global feature representation. In this part we deal with latent representation manipulation and how in this case it influence the decoding part. varying the parameters of this part in fact is possible to notice how these variation can lead to different results from a PCP viewpoint and more robustness to far subjects. in fact one in particular of our test has been variate the subject of the image from a very near subjects to the camera to a more far representation given by different dataset used for Human pose estimation from UAV. It is possible to see that the results in case of a simple person in front on the camera are superior but when the limbs of the subject are composed by very small groups of pixel a more pixel-by-pixel analysis is needed. In terms of result we obtained a level of PCP for 50 epoch-training showing the challenging properties of the images. The parameters modified are in fact not only the kernel sizes but also the dilation or rate parameters obtaining in this way a general FoV of the image. We also tried to use different sizes for the Latent representation, needed to use the dilation high and use bigger FoV than we can and more parallel levels of FoV concatenated for the decoding part. The shapes of the layers variate between 1,2,6 rate parameters while the dimension of the kernels between 3,5,7. Another important fact is related to the presence of the 1 by 1 convolutions before the concatenation useful if we want to manipulate the dimension of the data without loss of features from the local to the global feature map.

4. The DWT-based Encoder

The methodologies applied for the construction of the information related the analysis of the image in frequency with different scale lead me to many different choice, from the application of Gaussian or Sobel Filters to the use of SRM structure as in [22]. But what in the end establish the most significant result has been the DWT encoder with the multi-scale decomposition of the image. To make it we built a sequence of layer applying low pass and high pass filters to an image and generating relevant Haar-features for the localization of relevant key points.The Multilevel decomposition method called in this work DWT-based encoder will generate these coefficients over all three direction in the image vertical, diagonal and horizontal (LH_i, HL_i, HH_i, LL_i) with coefficients for the details and approximation over different thresholds to pass to the next layer as explained in [23],[24] or [25]. With a more mathematical viewpoint each frequency component can be defined in a matrix form for 2D input as:

$$\chi_{ll} = \Gamma \chi \Gamma^T \tag{2}$$

$$\chi_{lh} = \Gamma\chi K^T \tag{3}$$

$$\chi_{hl} = K\chi\Gamma^T \tag{4}$$

$$\chi_{hh} = K\chi K^T \tag{5}$$

Where χ is the input, χ_{ll} are the low frequency component and the high defined with $\chi_{hl}, \chi_{lh}, \chi_{hh}$ Defining the Low pass and High pass filter as $\kappa i, \gamma_i$:

$$\Gamma = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots \\ \gamma_{-2} & \gamma_{-1} & \gamma_0 & \dots & \dots \\ \gamma_{-1} & \gamma_0 & \gamma_1 & \dots & \dots \\ \dots & \gamma_1 & \gamma_2 & \gamma_3 & \dots \\ \dots & \dots & \gamma_1 & \gamma_2 & \dots \end{pmatrix}, \; \mathrm{K} = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots \\ \kappa_{-2} & \kappa_{-1} & \kappa_0 & \dots & \dots \\ \kappa_{-1} & \kappa_0 & \kappa_1 & \dots & \dots \\ \dots & \kappa_1 & \kappa_2 & \kappa_3 & \dots \\ \dots & \dots & \kappa_1 & \kappa_2 & \dots \end{pmatrix} \tag{6}$$

These information produced for each layer will be concatenated to the layers of the encoder but, having a U-net structure will be added also at the last layers with forward propagation that will be taken into account by the convolution's weights and updated by back-propagation:

$$\frac{\partial\chi_{ll}}{\partial\chi} = \Gamma^T G\Gamma \tag{7}$$

$$\frac{\partial\chi_{hl}}{\partial\chi} = \Gamma^T GK \tag{8}$$

$$\frac{\partial\chi_{lh}}{\partial\chi} = K^T G\Gamma \tag{9}$$

$$\frac{\partial\chi_{hh}}{\partial\chi} = K^T GK \tag{10}$$

This is used in the encoder But not as a down-sampling operation to substitute in the encoder instead its down-sampled version is added hierarchically to the layers as a Parallel encoder providing in this way to the three encoders, but also decoder considering that the structure is U-net as, the features stressed by each layer of the multi-level decomposition. As it is possible to see the application of different low pass filters applied for the first,fifth and ninth image for each layer and high pass filter for the rest. These information concatenated will be added hierarchically to the encoder layers in particular in the 2-th,3-th and 4-th, each level will divide the dimension of the image with 2^j with j=number of the layer. In order to recap how wavelet works I'm referring to different mother wavelet as Haar wavelets but we will also analyze performances in correspondence of different wavelets applied as Daubechies that already proved their capabilities in the isolation of high from low frequency components in images and isolation in all directions of edges at different scale and resolutions as in [2] and [24]. Another remarkable fact is that we do not need to use IDWT in decoding for obvious reason and the fact that, having a simple Multi level decomposition without variation filters. We will have just one gradient in common to consider for the loss minimization evaluating the additional DWT features directly in the same CNN's loss as in the case of [15] with different loss for each heat map to find for each key point an unique connection to the others for the skeleton construction. our approach is based on the analysis of the information produced by this filters and the improvement given by the analysis of the image by the DWT encoder as substitute to the one based on SRM. In order to better understand the uses of DWT in this part is given a recap of basics concepts. Given a window function as the one used for Fourier transform usually found in a common form:

$$F(\tau,\epsilon) = \int_{-\infty}^{+\infty} f(t)g(t-\tau)e^{-it\epsilon}dt \tag{11}$$

It can be interpreted as a Fourier transform of f at the frequency ϵ, localized by the window g in the neighborhood of τ. Multiplying the signal represented by f(t) with g and computing the Fourier coefficients we obtain indication of the frequency content of the signal f in a neighborhood of τ, shifting the window from 0 and obtained a sequence of coefficients that give a representation of the image sensible to certain frequencies. Now, considering g as the family of function generated from a single $L^2(R)$ function by phase space translations (τ,ϵ) where ϵ = 1/s ("coherent states"), an important property of this function is the capability to completely reconstruct f from the phase space projections given by $\langle g^{(\tau,q)}, f\rangle$. This is due to the property of this mapping function of being an isometry that as mentioned in [24],[26] or [27] is given by so called resolution of the identity property that implies that the f function can be written as:

$$f = \frac{1}{2\pi}\int d\tau \int dq \langle g^{(\epsilon,\tau)}, f\rangle g^{(q,\tau)} \tag{12}$$

In similar way the wavelets are family of functions that involve the $g^{(\tau,q)}$ derived from a function, but indexed by two labels, one for position and one for frequency with s = $1/\epsilon$ as scale factor and τ = translation where the resolution of the identity is written as:

$$f = C_\psi^{-1}\int \frac{dq}{q^2}\int d\tau \langle \psi^{(\tau,q)}, f\rangle \psi^{(\tau,q)} \tag{13}$$

Taking into account the (14) in this way we can redefine completely f with a set of coefficients over a direction generated by simple filter application and re-defining the image using:

$$F(\tau,s) = \frac{1}{\sqrt{|s|}}\sum_1^{p-1} f(t)\psi^k[\frac{t-\tau}{s}] \tag{14}$$

Usually it is chosen as parameters $s = 2^j$ as dilation in order to have a discrete dilation by taking powers of a fixed j, $\tau = 2^j n$ as translation of the wavelet and k as direction rewriting $g^{(\tau,q)}$ as $\psi^k_{j,n}$. Given a function $f(t)$ as signal of input that will be our image, it has a large amplitude near sharp transitions of pixels such as

edges, obtaining the coefficients over the $\langle \psi^k_{j,n}, f \rangle \geq T$ threshold and varying it over the frequency. What now is produced by this are three corresponding high pass and low pass filters we obtain four results As approximation and details coefficients for each of the three layers to pass to the next one in the multi level decomposition where each of theme is defined as:

$$\begin{cases} A^d_{2^j} f = (f \gamma_{2^j}(-x)\gamma_{2^j}(-y))(2^{-j}n, 2^{-j}m) \\ D^1_{2^j} f = (f \gamma_{2^j}(-x)\rfloor 2^j(-y))(2^{-j}n, 2^{-j}m) \\ D^2_{2^j} f = (f\lfloor 2^j(-x)\gamma_{2^j}(-y))(2^{-j}n, 2^{-j}m) \\ D^3_{2^j} f = (f\lfloor 2^j(-x)\rfloor 2^j(-y))(2^{-j}n, 2^{-j}m) \end{cases} \quad (15)$$

Another important remark is on how the output is associate and added to the neural network this is related the method of concatenation (fusion) and the corresponding result. In case of application of hierarchical fusion has been proved an increment booth in velocity of the loss convergence and PCP metric evaluated.

5. Experimental Setup

In our case is necessary to give some hint for the evaluation of performances. we used the commonly used method of Percentage of Correct Parts (PCP) and Percentage of Correct Key-points (PKC) to evaluate the results. In particular considering a simple MSE function to minimize the error respect predicted and correct coordinate we have to consider a simple euclidean distance to minimize (as in a regression problem) in the simple form:

$$d^2 = (x_{predicted} - x_{GT})^2 + (y_{predicted} - y_{GT})^2 \quad (16)$$

Having this kind of loss we have the possibility to define a suitable metric depending on these coordinates for the evaluation based on a threshold to apply in order to understand if I'm going near to desired result. The PKC will be denoted an evaluation of the joints by a formula in the form:

$$d^2_{pred} cos(\theta) \leq 0.5 d^2_{true} \quad (17)$$

In other words if the segment given by the predicted endpoints lie within fraction of the length of the ground-truth segment the distance calculated by the prediction will have to be smaller than the half of the effective lenght (threshold = 0.5) as mentioned in [1]. Alternatively is possible to use as metric the Object Keypoint Similarity (OKS) in the form:

$$\frac{\sum \exp d_i^2/2s^2k_i^2 f(v_i > 0)}{\sum f(v_i > 0)} \quad (18)$$

Another possible metric the one adopted in this work is the PCP that is based on detected joint that is considered correct if the distance between the predicted and the true joint is within a certain threshold. In this work you can find an example of the results in terms of accuracy

$_{params}{}^{\diagdown}{}_{\diagup}{}^{tests}$	**Variations**	**diff. score**
output feature	(36×36)	-
input images	(256×256)	-
Optimizer	Adam/SDG	+0.17 loss/epoch
Activation function	Lrelu/Relu	-
Num.layer	3/4	$\pm 0.1 pcp$
N.layer(loss)	1/3	-0.08 loss/epoch
Batch	16/32	+0.2 % pcp
Learning rate	$3e^{-3}$/0.0025	+0.2% loss/epoch
Epochs	10/30/120	+0.11→+0.5% pcp
Dropout	0.2/0.5	3.0% loss/epoch
Wavelet	Haar/Daub.	+0.05→+0.14% pcp

Table 1
Parameters used in the tests

PCP on a single image of the adopted methods varying during the training. It is relevant to consider that we have to save many coefficients in the computation of the wavelet so the computation will be calculated before the training function. In this way we will gain some time in the computation repeated for each epoch. But this is now a very challenging problem from the memory point of view because we have to occupy space for each image and the corresponding coefficients.

6. Results

The tests are evaluated both for COCO and LSP datasets but in the end, the evaluation has been done on a combination of it.The results in accuracy are evaluated not from the first epoch but from the 50th epoch while the loss is shown from the beginning. It is important to mention that the approaches have been proven on 160x160 images but also different dimensions and increasing them the results increased also more with the DWT encoder structure with respect to a simple CNN just proving that the loss, over 100 epoch arrived from a 110.07 value to 120.3. Not a big increase but considering that the PCP arrived from a value of 65% to 71% and that the loss of an interpolation problem for the 160x160 images start usually over 1000 as MSE loss initial value while 300 for 80x80 images we can deduce that the augmented complexity of the interpolation problem is compensated from the information provided from DWT confirming its utilities for the analysis of complex data, It could be interesting to try with 1280x720 images as a future improvement.

Some of the most challenging aspects respect this kind of problem are reported in the 4-th image at the top that are evaluated with our method confronted at parity of epoch training and parameter to a simple CNN with WASP without addition of DWT. better results can be establish, without changing the parameter, eliminating the down sampling and using images that contains a better resolution (instead a 160x160 image) as in [15] where image's dimensions are around 1280x720 even with de-

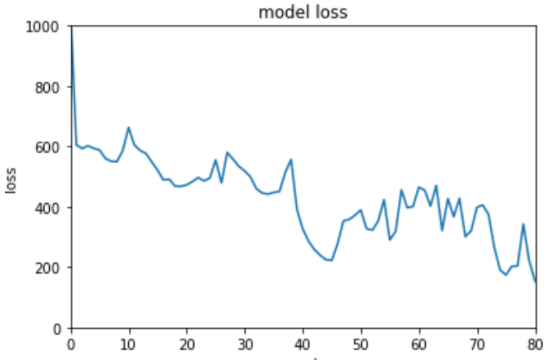

Figure 1: Behaviour of the loss with 80 epoch with WASP bottleneck on the dataset LSPII calculated over smaller images 80x80 and on COCO dataset, what is clearly visible beyond the lower loss is the irregularity of the function in this case due to noised data that is the main difference from LSPII and COCO subset selected by me and the lower resolution of the images.

nonised images. The images have been chosen to give a basic representation of the main problems of the human body pose estimation due to the complexity of the pose and occlusion of the limbs of the subject. Looking at the first line the first picture, beyond a small negligible error, denote a very good result in the pose estimation of the subject. Differently, the third image is more complex to evaluate due to the occlusion of the limbs and the complexity of the pose. A different kind of problem is instead represented by the pose of the second and third image where the entire image down sampled with a very high factor leads to the problem of low border resolution of the subject giving imprecision in the evaluation of the key points positions. Note carefully the presence of padding in the CNN leads us to the shifted results in the figure as in object detection as [28]. In an other important test we considered the DWT based method confronted between COCO and LSPII dataset where the data are more problematic and many subject assume complex poses. These tests are also evaluated with respect a PKC value but in the end we used the PCP because more appropriate for a bottom-up approach and easier to implement and evaluate but similar results are initially evaluated with respect the PKC.

7. Conclusion

The wavelet procedure actually increase the capabilities with the DWT encoder and with this result how is possible to extend it also to different fields that include these kind of structures. Different results are obtained using more epochs or layer in the convolution structure from 4 to 3 layers obtained lower results in PCP terms so more

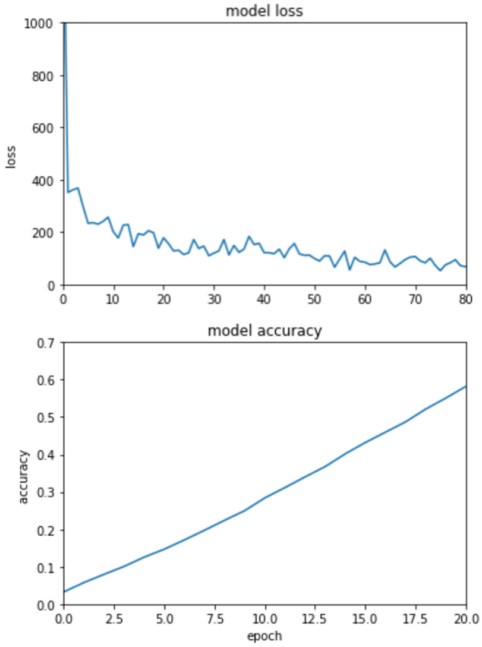

Figure 2: Test with HaarDWT used with pretarined CNN Daubachies.

complicated CNNs are not capable to compete with the wavelet encoder showing clear results of over fitting so we preferred to do not modify epochs more than 100 and using just 4 convolution layers. We extended these result even to a different topic as human pose estimation, the initial objective was to establish results in object detection approach for pose estimation but it has been discard because as already said the top-down approach proved to be superior. In addition to this we solved the faulty encoder-decoder general structure common to all most used CNNs for this field as VGG-16 and ResNet showing lower loss of information during encoding using the DWT and the analysis of the Wavelet's information of the images.

References

[1] M. Eichner, M. Marin-Jimenez, A. Zisserman, V. Ferrari, 2D articulated human pose estimation and retrieval in (almost) unconstrained still images, International Journal of Computer Vision 99 (2012) 190–214.

[2] J. C. Isaacs, S. Y. Foo, Hand pose estimation for american sign language recognition, Thirty-Sixth Southeastern Symposium on System Theory, 2004. Proceedings of the (2004) 132–136.

[3] S. Pepe, S. Tedeschi, N. Brandizzi, S. Russo, L. Iocchi, C. Napoli, Human attention assessment us-

Figure 3: Examples of qualitative result of our model including some incorrect classification in the image 3 where the PCP is very law note how the most problematic one are the first and second in the parts where we have limb's occlusion.

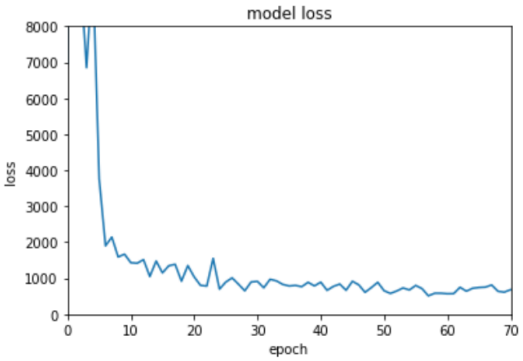

Figure 4: Another test with different kernel's sizes in the WASP that obtained better results but with padding that decreased the result's PCP (see the colab).

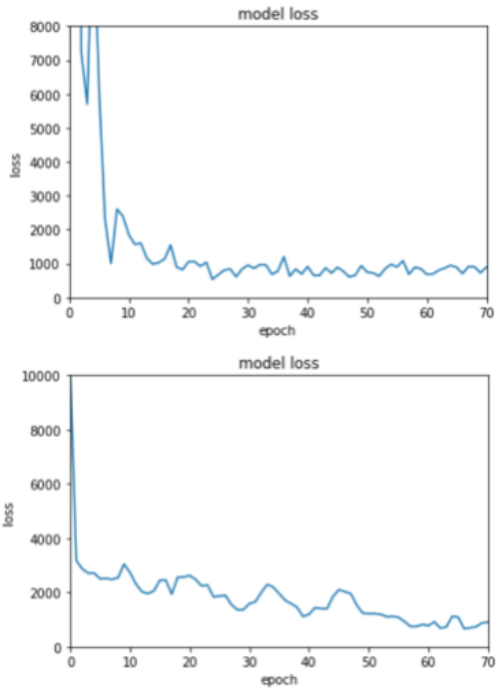

Figure 5: Compared final results over the epoch as result of DWT+WASP and Daubechies compared between LSPII+COCO and COCO only dataset.

Method	COCO	LSPII
$CNN_{wasp+DWT,haar}$	76.45%	71%
$CNN_{DWT,haar}$	75.02%	-%
CNN_{wasp}	77%	63%
CNN$_{wasp+DWT,Daubachies}$	74.1%	71%
U-net	73.7%	69.3%
$CNN_{wasp,SRM}$	55.33%	60.21%
$CNN_{DWT,Daubachies}$	61%	-%

Table 2
The results with respect a PCP metric evaluated over 100 epoch, a batch of 40 elements each tested on LSPII dataset and over COCO for U-net for the CNN with different structures. The - values are not interesting with respect the previous result in the table (e.g same % for U-net and $CNN_{DWT,haar}$ or $CNN_{DWT,haar}$ and $CNN_{DWT,Daubachies}$ in COCO).

ing a machine learning approach with gan-based data augmentation technique trained using a custom dataset, OBM Neurobiology 6 (2022). doi:10.21926/obm.neurobiol.2204139.

[4] N. Dat, V. Ponzi, S. Russo, F. Vincelli, Supporting impaired people with a following robotic assistant by means of end-to-end visual target navigation and reinforcement learning approaches, in: CEUR Workshop Proceedings, volume 3118, CEUR-WS, 2021, pp. 51–63.

[5] F. Bonanno, G. Capizzi, G. Sciuto, C. Napoli, G. Pappalardo, E. Tramontana, A novel cloud-distributed toolbox for optimal energy dispatch management from renewables in igss by using wrnn predictors and gpu parallel solutions, in: 2014 International Symposium on Power Electronics, Electrical Drives, Automation and Motion, SPEEDAM 2014, IEEE Computer Society, 2014, pp. 1077–1084. doi:10.1109/SPEEDAM.2014.6872127.

[6] C. Napoli, G. Pappalardo, E. Tramontana, R. Nowicki, J. Starczewski, M. Woźniak, Toward work groups classification based on probabilistic neural network approach, in: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), volume 9119, Springer Verlag, 2015, pp.

79–89. doi:10.1007/978-3-319-19324-3_8.

[7] Q. Li, L. Shen, S. Guo, Z. Lai, Wavelet integrated cnns for noise-robust image classification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[8] M. Wozniak, C. Napoli, E. Tramontana, G. Capizzi, G. Lo Sciuto, R. Nowicki, J. Starczewski, A multiscale image compressor with rbfnn and discrete wavelet decomposition, in: Proceedings of the International Joint Conference on Neural Networks, volume 2015-September, Institute of Electrical and Electronics Engineers Inc., 2015. doi:10.1109/IJCNN.2015.7280461.

[9] N. Brandizzi, V. Bianco, G. Castro, S. Russo, A. Wajda, Automatic rgb inference based on facial emotion recognition, in: CEUR Workshop Proceedings, volume 3092, CEUR-WS, 2021, pp. 66–74.

[10] C. Napoli, G. Pappalardo, E. Tramontana, Using modularity metrics to assist move method refactoring of large systems, in: Proceedings - 2013 7th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS 2013, 2013, pp. 529–534. doi:10.1109/CISIS.2013.96.

[11] G. Capizzi, G. Sciuto, C. Napoli, E. Tramontana, A multithread nested neural network architecture to model surface plasmon polaritons propagation, Micromachines 7 (2016). doi:10.3390/mi7070110.

[12] G. De Magistris, R. Caprari, G. Castro, S. Russo, L. Iocchi, D. Nardi, C. Napoli, Vision-based holistic scene understanding for context-aware human-robot interaction 13196 LNAI (2022) 310–325. doi:10.1007/978-3-031-08421-8_21.

[13] R. Brociek, G. Magistris, F. Cardia, F. Coppa, S. Russo, Contagion prevention of covid-19 by means of touch detection for retail stores, in: CEUR Workshop Proceedings, volume 3092, CEUR-WS, 2021, pp. 89–94.

[14] R. Avanzato, F. Beritelli, M. Russo, S. Russo, M. Vaccaro, Yolov3-based mask and face recognition algorithm for individual protection applications, in: CEUR Workshop Proceedings, volume 2768, CEUR-WS, 2020, pp. 41–45.

[15] B. Artacho, A. Savakis, Unipose: Unified human pose estimation in single images and videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[16] J. Dong, W. Jiang, Q. Huang, H. Bao, X. Zhou, Fast and robust multi-person 3d pose estimation from multiple views, 2019. arXiv:1901.04111.

[17] G. Capizzi, C. Napoli, F. Bonanno, Innovative second-generation wavelets construction with recurrent neural networks for solar radiation forecasting, IEEE Transactions on Neural Networks and Learning Systems 23 (2012) 1805–1815. doi:10.1109/TNNLS.2012.2216546.

[18] B. Artacho, A. Savakis, Omnipose: A multi-scale framework for multi-person pose estimation, 2021. arXiv:2103.10180.

[19] Y. Zhou, W. Huang, P. Dong, Y. Xia, S. Wang, D-unet: A dimension-fusion u shape network for chronic stroke lesion segmentation, IEEE/ACM Transactions on Computational Biology and Bioinformatics 18 (2021) 940–950. doi:10.1109/TCBB.2019.2939522.

[20] T. Williams, R. Li, Wavelet pooling for convolutional neural networks, 2018.

[21] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, 2015. arXiv:1411.4038.

[22] S. Kang, H. Park, J.-I. Park, Cnn-based ternary classification for image steganalysis, Electronics 8 (2019). URL: https://www.mdpi.com/2079-9292/8/11/1225. doi:10.3390/electronics8111225.

[23] H. Liu, N. Chen, J. Huang, X. Huang, Y. Q. Shi, A robust dwt-based video watermarking algorithm, 2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No.02CH37353) 3 (2002) III–III.

[24] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Transactions on Information Theory 36 (1990) 961–1005. doi:10.1109/18.57199.

[25] B. Sturm, Stéphane mallat: A wavelet tour of signal processing, 2nd edition, Computer Music Journal - COMPUT MUSIC J 31 (2007) 83–85. doi:10.1162/comj.2007.31.3.83.

[26] I. Daubechies, T. Paul, Time-frequency localisation operators-a geometric phase space approach: Ii. the use of dilations, Inverse Problems 4 (1988) 661–680.

[27] S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (1989) 674–693. doi:10.1109/34.192463.

[28] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, Lecture Notes in Computer Science (2014) 346–361. URL: http://dx.doi.org/10.1007/978-3-319-10578-9_23. doi:10.1007/978-3-319-10578-9_23.