https://doi.org/10.52326/ic-ecco.2022/CS.11

Human Motion Recognition Using Artificial Intelligence Techniques

Andrei Enachi [1], ORCID: 0000-0002-1000-457X
Cornel Turcu [2], ORCID: 0000-0002-3134-8769
George Culea [1], ORCID: 0000-0001-8286-1602
Ioan-Viorel Banu [1], ORCID: 0000-0002-7722-8017
Dragos-Alexandru Andrioaia [1], ORCID: 0000-0002-0244-7688
Puiu-Gabriel Petru [1], ORCID: 0000-0003-2864-3422
Sorin-Eugen Popa [1], ORCID: 0000-0002-1939-587X
[1] "Vasile Alecsandri" University of Bacău, Calea Mărășești 157, Bacău, 600115, Romania
[2] "Stefan cel Mare" University of Suceava, 13, University Street, Suceava, 720225, Romania
andrei.enachi@ub.ro

***Abstract*— The goal of this paper's research is to develop learning methods that promote the automatic analysis and interpretation of human and mime-gestural movement from various perspectives and using various data sources images, video, depth, mocap data, audio, and inertial sensors, for example. Deep neural models are used as well as supervised classification and semi-supervised feature learning modeling temporal dependencies, and their effectiveness in a set of tasks that are fundamental, such as detection, classification, and parameter estimation, is demonstrated as well as user verification.**

A method for identifying and classifying human actions and gestures based on utilizing multi-dimensional and multi-modal deep learning from visual signals (for example, live stream, depth, and motion - based data). A training strategy that uses, first, individual modalities must be carefully initialized, followed by gradual fusion (called ModDrop) to learn correlations between modalities while preserving the uniqueness of each modality specific representation. In addition, the suggested ModDrop training approach assures that the classifier detect has weak inputs for one or maybe more channels, enabling these to make valid predictions from any amount of data points accessible modalities. In this paper, inertial sensors (such as accelerometers and gyroscopes) embedded in mobile devices collect data are also used.

Keywords*— *learning methods; ModDrop; neural models; sensors; mime-gesture

I. INTRODUCTION

Computer vision is a field of research that deals with the analysis of human motion from visual input. Over the years, this has been a daunting task, but much progress has been made. The problem of obesity is a focus for many people, as it is a serious health issue. Hundreds of applications that could benefit from this technology currently exist, including applications that control navigation and manipulation in real and virtual environments, interaction between humans and robots, telepresence systems, and more, portals that assist the hearing and speech impaired, learning tools and games, engineering systems as well as computer-aided design, as well as automatic video annotation and indexing are all geared towards aiding those with hearing and speech disabilities. Forensic identification and lie detection are other applications that may be helpful to those with impaired senses [1] – [5].

Different research communities have approached application-driven research in human motion analysis from various perspectives. Today, facial expression interpretation and crowd movement analysis have become more standardized tasks. There are many taxonomies However, most existing work in this domain focuses on group or individual activities, primitive actions or gestures. Each problem can be thought of as a variant recognition or reconstruction, one of two classic computer vision challenges [2], [3].

The gesture and myth, the hand movement accompanied by the speech, can not only provide meaningful meaning, but also provide information about the personality and cultural differences of the speaker, immediate emotional state, intentions and attitudes toward the audience, and the topics ahead of the eye. Some psychological studies suggest that motor movements are not only used to illustrate spoken words, amplify their emotional impact, or compensate for a lack of linguistic

fluency, but in turn will influence the process of speech production. In addition to verbal information, the ability to perceive and comprehend nonverbal clues would be advantageous to make human-machine communication easier and more intuitive regarding the interpersonal differences between the subjects or gestures, here the object of distinction is changed so that the individual is recognized regardless of the tasks or gestures used or performed [3], [4].

II. VISUAL RECOGNITION

In this section we will look at traditional training techniques and strategies for recognizing human movements and mime-gestures from visual data, using static descriptors of a frame, and last but not least dynamic ones, together with their anatomical links.

Although most of the discussed manuscripts are unrelated in relation to the research presented in this article (in terms of the methodology used, deep learning) but still provide the background and insights necessary to understand the challenges of the presented field here. For more general methods that can help inspire our work, the scope of the analysis of human movement and gesture is very broad, covering several tasks performed simultaneously both temporally and spatially (through recognition of mimic and gesture, action and activity performed). Similar problems can be used to solve other problems, with a significant role, being abstracted according to level.

The problems to be studied in here will focus on the action and/or activity of a person, in a supervised manner intervening only where appropriate. Identity, gesture and mimic recognition (as well as sign interpretation), is done by introducing a visual input as a category within the sign language or existing signs. Learning can be done through videos, but also dimensionally from a third-person perspective.

The research area(s), covering hand motion modelling and analysis, can then be fur quantified into a set of typical problems, listed below (see Fig. 1): identification of the person from gestures, the reverse of the previous stage, and no need to classify the person performing a gesture into a category. Another aspect of this research, associated from the same group, is the recognition of the whole body and not just the position and/or hand, gait; approximation of the hand position, from a mapped dimensional point for high visualization of the possible hand positions a person performs, taking into account the person's anatomical constraints.

Figure 1. Detection of human movements: mimic-gesture identification, mocap, position from which the person was captured understanding and plausibly simulating human behaviour

Hand position estimation can also be extrapolated within the body; in order to train the algorithm from the hands, will be by optimizing the optics and creating specific conditions to achieve high accuracy (thus achieving accuracy on the stored data). For capturing the movements performed by the body, it is not difficult to capture; the interpretation of the hand-object position, being involved its recognition as well as the remodelling for the gestures and mimicry performed, but also the issues concerning the interaction with the object(s).

Traditional methods in following steps are typically used for action recognition and location estimation: the pre-processing operations are those that are carried out before the data is analyzed. This includes things like cleaning up the data, removing errors, and formatting it correctly, locating and assembling or acquiring knowledge visualizations (for example, measuring, background subtraction, segmentation, and noise). As a cascade of following data reformulations and compressions, this procedure can be performed in a sequence of phases. Classification, involving one or more machine learning algorithms; or optimization based on a predefined metric. Deep learning techniques, as was described at the beginning of this chapter [6].

III. PREPROCESSING AND MANUAL DETECTION

The preliminary steps consist of a collection of well-known depending on the type of input data, techniques and routines are used in various combinations. The goal of this step is to locate and segment the object of interest. (in our case, hands) in the background, eventually, noise and other artifacts will be compensated. Some methods and datasets make assumptions that this step is done ahead of time; however, in practice, the quality of preprocessing data is an issue that can have a significant impact on the performance of an algorithm. Various approaches to hand localization that rely on color images frequently include skin color detection. It works well in simple contexts because skin tones work well in appropriately chosen color spaces (see Fig. 2).

Pixel figures are created by combining the corresponding colors. Several studies have been conducted on color spaces and skin detection methods. HSI, in particular, has been shown to be one of the best color spaces for this task. Hand detection, on the other hand, which is solely based on skin color, remains extremely sensitive to illumination, ethnic divergence, and individual and background differences.

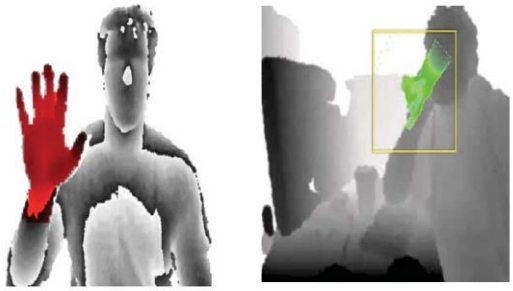

Figure 2. Example of pixel-based hand detection (left) and a case of failed hand position estimation in a poorly segmented hand case (right). [6]

A more robust face detector is frequently used as a practical ad hoc solution first to compensate for illumination differences and estimate the expected skin tint, as well as to define an approximate expected hand position. By training an SVM on extracted HoG descriptors, for instance, it is possible to learn a hand detector from a large amount of data. Van den Bergh and van Gool proposed utilizing a post Gaussian Mixture Model (GMM) to identify distinct skin hues in varied light levels. The GMM is combined with skin color estimation from a given face of each user in this approach, and the resulting probability that each pixel belongs to a skin region is then reported to a threshold value to produce a model.

They also improve their method by calculating and applying a face-to-hand distance priority in three dimensions utilizing depth data from a time-of-flight camera with the advancement of range sensors, the number of dimensions available is the number of spatial dimensions and color channels available for extracting spatial features has been increased to 6 (or, technically accurate, 5.5). But, most present systems don't really consider these equally. For example, color pictures are generally utilized solely for hand detection. Color images, for example, are frequently used exclusively for hand detection, whereas depth maps are used to compute discriminative features for the following gesture classification. Simultaneously, a number of papers use the opposite strategy, in which hand positions are computed using color or intensity descriptors are used in point cloud segmentation and spatial features.

IV. CONCLUSIONS

The primary goal of this work is to develop the most advanced automatic methods for analyzing and interpreting human and mime-gesture movements, as well as sign language interpretation, from various perspectives and using various data sources such as images, videos, depth maps, mocap analysis data, audio, and inertial sensors. In the end, is proposed a set of deep neural network models with specific algorithms for classification by surveillance and semi-surveillance for learning, for modeling time and spatial dependencies, and for achieving efficiency and effectiveness on a set of fundamental tasks such as detection, classification, and user identity verification parameters.

Here is presented a method for recognizing human and mimic-gesture activities, classification-based fusion of multiple neural network models, and visual data learning. The training strategy that will be the foundation of this work, as it will be the fusion of deep neural models with separate data channels (ModDrop) to achieve intersecting model learning while preserving the uniqueness and specific mode of data representation.

Moving away from 1-to-n mapping and toward continuous evaluation of user gesture and mimicry parameters, the hand position problem, and a new method for the regression step within the image depth map using deep convolutional neural networks, where imprecise, incomplete data are merged into an intermediate representation of the hand in segmented form are all addressed. In papers related to this topic, we investigate deep convolutional neural network models for identifying users based on their movements. The data that will be stored based on embedded inertial sensors in cameras and/or mobile devices.

REFERENCES

[1] A. Agarwal and B. Triggs, “Recovering 3D human pose from monocular image,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.1, pp. 44–58, 2006.

[2] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, vol.43, no. 3, pp. 1–43, 2011.

[3] A. Aldoma, M. Vincze, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu, and G. R. Bradski, “CAD-model recognition and 6DOF pose estimation using 3D cues,” *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 585-592, 2011.

[4] L. A. Alexandre, A. C. Campilho, and M. Kamel, “On combining classifiers using sum and product rules,” *Pattern Recognition Letters*, vol. 22, no. 12, pp. 1283–1289, 2001.

[5] D. Amodei, R. Anubhai, Case C. Casper J. Battenberg, et al., “Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin,” Preprint, arXiv:1512.02595, 2015.

[6] Costin Anton Boiangiu, Marius Eduard Cojocea, Robert Costin Bercaru, Mihai Bran, Mihai Lucian Voncila, Nicolae Tarba, Cornel Popescu, George Culea, “*Fast and Reliable Emotions Detection Adapted for Driver Monitoring and Online Psychotherapy Sessions*, *CEAI*, vol. 24, no. 3, 2022.