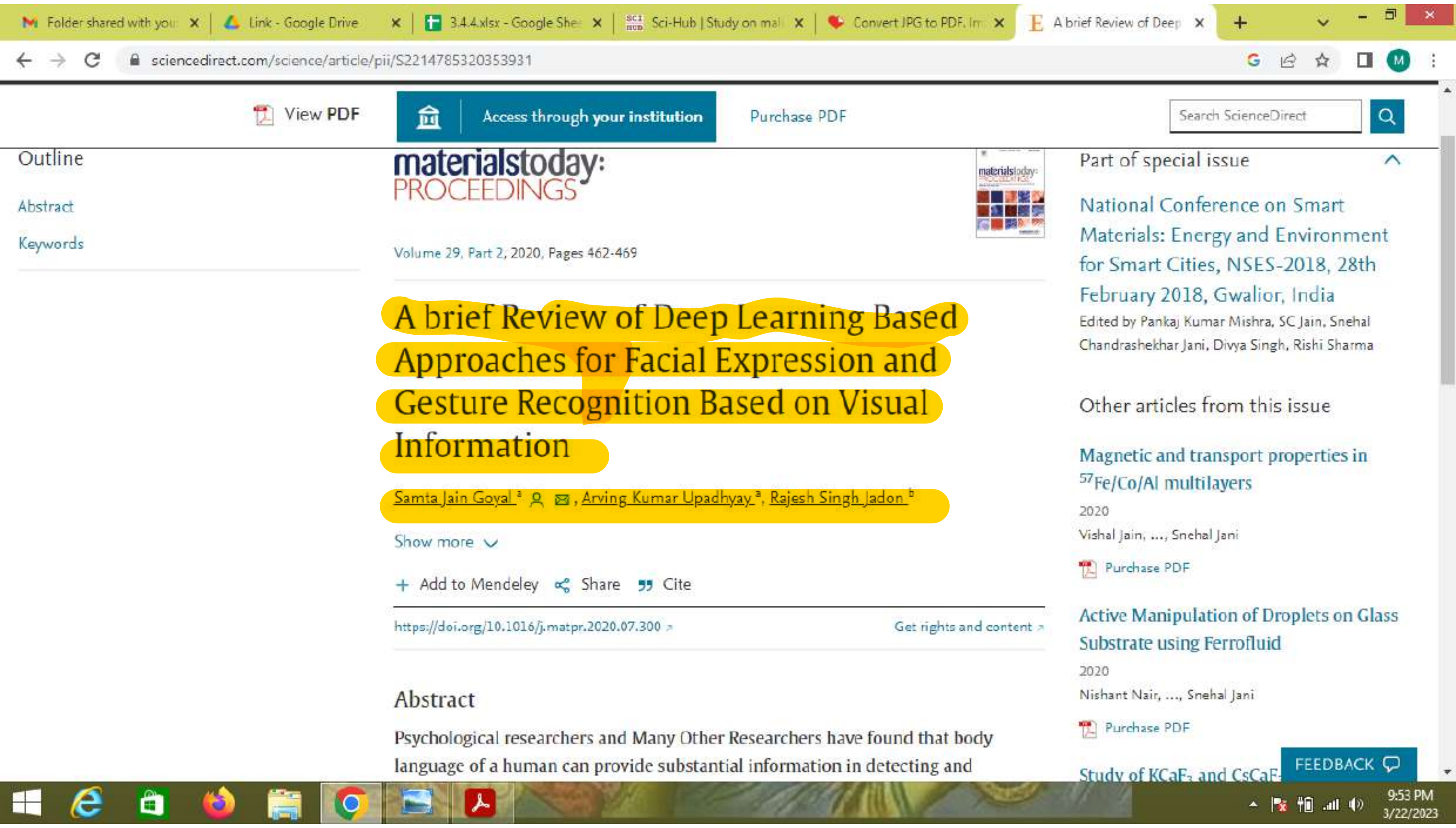


Volume 29, Part 2, 2020, Pages 462-469

A brief Review of Deep Learning Based Approaches for Facial Expression and Gesture Recognition Based on Visual Information

Samta Jain Goyal [a], Arving Kumar Upadhyay [a], Rajesh Singh Jadon [b]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.matpr.2020.07.300

Get rights and content

Abstract

Psychological researchers and Many Other Researchers have found that body language of a human can provide substantial information in detecting and

Part of special issue

National Conference on Smart Materials: Energy and Environment for Smart Cities, NSES-2018, 28th February 2018, Gwalior, India

Edited by Pankaj Kumar Mishra, SC Jain, Snehal Chandrashekhar Jani, Divya Singh, Rishi Sharma

Other articles from this issue

Magnetic and transport properties in ^{57}Fe/Co/Al multilayers

2020

Vishal Jain, ..., Snehal Jani

Purchase PDF

Active Manipulation of Droplets on Glass Substrate using Ferrofluid

2020

Nishant Nair, ..., Snehal Jani

Purchase PDF

Study of KCaF$_3$ and CsCaF

Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings 29 (2020) 462–469

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

NSES 2018

A brief Review of Deep Learning Based Approaches for Facial Expression and Gesture Recognition Based on Visual Information

Samta Jain Goyal[a], Arving Kumar Upadhyay[a], Rajesh Singh Jadon[b]

Amity School of Engineering and Technology, Amity University Madhya Pradesh, Maharajpura Dang, Gwalior (MP)-474005

[b]Department of Computer Applications, MITS, Gwalior, India (MP) 474001

Abstract

Psychological researchers and Many Other Researchers have found that body language of a human can provide substantial information in detecting and interpreting emotions. It could express explicit and implicit information mutually of one's emotional state and intentions over multi-channel modalities. These imperative channels include eye gaze, head movement, facial expression, body posture and gesture and so on. This learning focuses on detecting emotional states from the body language of the hand and face using computer vision and soft computing techniques. The facial expression and gesture recognition have widely used and challenging task in the present scenario. This paper describes brief review of all approaches which are majorly categorized into two -categories where the first category belongs to Conventional approaches and other based on Deep-learning or Non - Conventional. This paper is basically used as a survey of deep-learning approaches for hand-gesture and Facial Expression Recognitions. We describe all review through the taxonomy of deep-learning approaches proposed architecture details, fusion strategies, datasets, way to treat temporal-dimensions of data, its main features, basic knowledge & general understanding of its challenges.

Selection and Peer-review under responsibility of National Conference on Smart Materials: Energy and Environment for Smart Cities (NSES-2018).

Keywords: FER, Conventional FER, Deep learning Based FER, CNN, Spatial features, Temporal features.

1. Introduction

Among human's nonverbal cues are essential in conveying interactive information which can be achieved through body language, while a just 7% of communication involved of genuine words [1]. Besides, human communication through body language could convey emotional state of a person such as depression, boredom, enjoyment, etc. Over

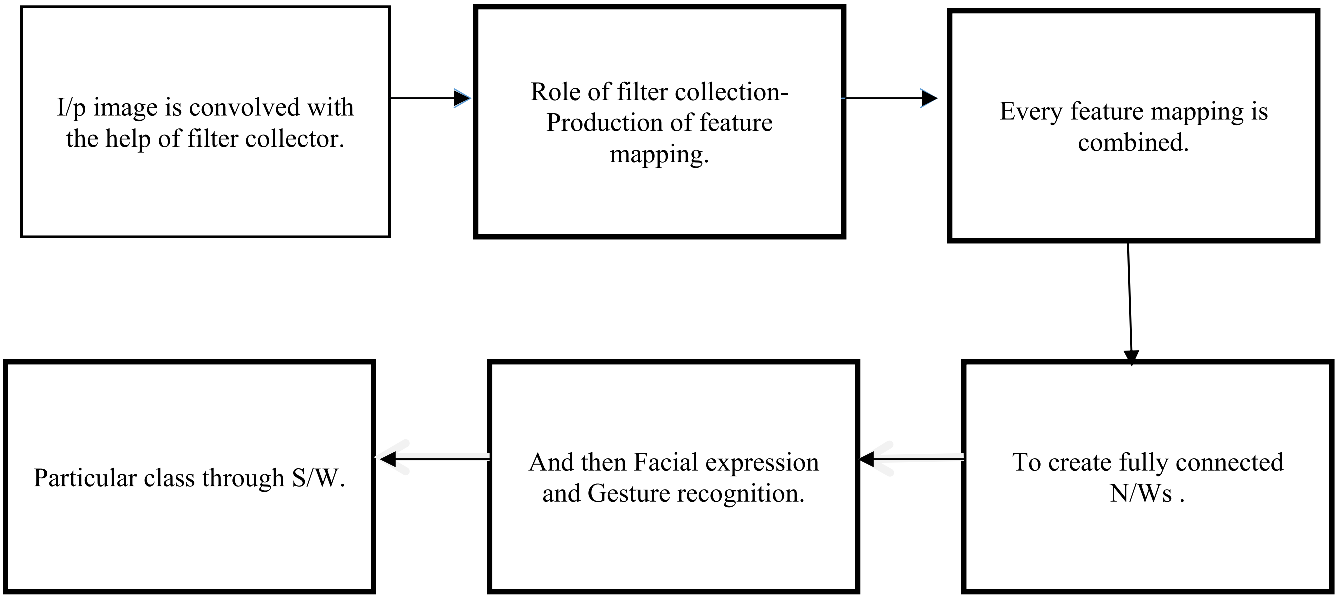
*Corresponding author. Tel.: +91998114220 *E-mail address:* sjgoyal@gwa.amity.edu

2214-7853 © 2018 Elsevier Ltd. All rights reserved.
Selection and Peer-review under responsibility of National Conference on Smart Materials: Energy and Environment for Smart Cities (NSES-2018).

body language of human, Verbal communication can easily be enhanced. For instance, if you understand the subject then u may present eye contact and nodding cues to the person that you are interacting with. The human brain is a multi-signal communicative media with significant tractability and specificity. When we communicate [2-3] we can gather essential information such as personal identification, ethnicity, gender, and age, attractiveness. Psychological researchers have found that body language of a human can provide substantial information in detecting and interpreting emotions [4-6]. Facial expression and gesture recognition are now primarily research in CV and AI field. These substantial progresses are going since last 2 decades. Facial emotion plays a major role in communication. To understand intensions, emotional states can be found based on facial expression, body gestures. Survey says that non -verbal communication is more effective than verbal. This type of study gaining lots of attention in the field of computer vision, AI and cognitive science including HCI, VR, AR, ADAS. Many times, various sensors can be used for recognition emotion and gesture.

This paper taking two segments and portions where 1st portion is for facial expression Recognition where all the approaches is divided into 2 -categories. One is known as conventional - FER-approaches and other based on deep learning. All conventional approaches have 3 major phases. First is face and facial component detection, feature extraction and at last classification of features. In deep - learning approaches, spatial temporal features are extracted from the components [7].

This paper objective is used to show the current trends for facial expression and gestures recognition. This paper analysis many papers in depth. In deep learning Network. Large amount of data and model complexity is a major challenge. RNN, LSTM are deep learning models used in image- sequence modelling in the area of facial expression Recognition and Gesture recognition. CNN is very important and more reliable network-model in deep -learning.


I/p image is convolved with the help of filter collector.
Role of filter collection- Production of feature mapping.
Every feature mapping is combined.
To create fully connected N/Ws .
And then Facial expression and Gesture recognition.
Particular class through S/W.

During analysis of many R-papers, it was found that dynamic Recognition rate is higher than Static Recognition.

1. Terminology

Some important and more frequently used terms are following in the field of Facial expression & Gesture Recognition: -

The facial action cooling system (FACS) is used to characterized facial action based on Facial-Emotions. These all are defined by [8,9]. Here facial action units (FAU) plays very important role, showing muscles movements &

Facial Landmarks (FL) are used to salient points of facial regions. Based on facial landmarks terminology, many approaches such as Active Shape-Based-Model (ASM), Appearance Based Model (AASM), Regression based model, CNN based approaches are used in current scenario.

2. Architecture

How to deal with temporal, Spatial dimensions or information, these are major challenges in deep- learning Based Recognition system. Here based on study, 3 categories are used for facial expression Recognition and Gesture Recognition. In the first category, approaches use 3D-filters in the convolutional layer which allow to capture discriminative features along with its spatial and temporal dimensions.

Second category uses Motion-Based-Features for pre-computation then extracted features are put into the N/W. Third category use individual or stacks of frames with a temporal modelling is applied. RNN is the best example of this category. It takes hidden layer in account. Basic issue with RNN network is of its short-term-memory, LSTM approach was used as a hidden layer of RNN [10]. Many useful and practical types of RNN is available in market such as B-RNN, H-RNN, D-RNN for the human expression and Gesture Recognition purpose. HMM which is a temporal modelling tool is also used for the same issues.

3. Concept of Fusion

In Deep- learning methods, fusion of information is required for Facial-Expression and Gesture Recognition from the segmentation phase, all the information of segmented components is combined or fused to get information from the trained model and from the multiple cues. Information fusion can be done through 3 ways in Deep-learning Models – First is Early infant fusion, second is late- fusion and last is combined approach for fusion information. Lots of variation based on different parameters are also approachable, especially in temporal – dimension. In fusion-strategies, ensembles or stacked networks come for the same issue [11-12].

4. Dataset

We explored many standard and relevant datasets based on Facial-Expressions and Gesture Recognition. Varieties of dataset are available with specific –functionalities. Generally, all efficient DB support combination of appearance feature as well as motion features because in appearance-based DB, it contains frame – level CNN-Description whereas Motion-Based approaches consider the top ranked parameter in their DB.

Generally, all standard DB consider the frames, its no., its modality, no. of classes. Some very popular & standard DB are UCF-101, THUMOS-14, JAFFE, MMI, CE, DISFA and [13]so on. All the standard dataset contains large number of samples. All the dataset has been used based on their comparative and extensive experiments nature. Traditionally, it contains 2D- Based-Dataset with their large samples, large pose variations and subtle behaviour of Facial and Gesture. Also, the analysis of 3D-image dataset is also a major challenge with the available dataset.

5. Challenges

There are some analysed challenges occurs in Computer- Vision-Field. Every dataset has some challenges or issues to fulfil the requirement of problems, generated by researchers. So difficult to fulfil the requirements. There we require some effective sensors, like cameras, Kinect sensors which are used to click pictures and images in various conditions like visible light, illumination conditions, natural and infrared light and so on. So thermal cameras, NIR cameras are required for this purpose.

6. Facial Expression & Gesture Recognition

a. Conventional Facial Expression and Gesture Recognition: -

There lots of research paper studied to get basics of conventional approaches for Facial-Expression Recognition. During study, it was found that in every conventional approach, first it detects the face region then extract geometric appearance and hybrid features of detected – face. [14] used two types of geometric feature based with 52-facial-landmark-points. There it first calculates the angle and Euclidean distance of each pair of landmarks of a frame is calculated. This is the first geometric feature of this research and in the second, distance and angle are subtracted from the corresponding distance and angles of the same frame. To classify the expression, multi-class- AdaBoost or SVM is used to create classifier [15-16].

Whereas to get the appearance features, it is extracted from the global-face-regions which contain different types of information [17,18]. Based on the global features, [19] uses Local-Binary-Pattern (LBP) histogram from a global face-region as feature-vectors and for classification of expression uses Principal-Component-Analysis (PCA). This approach faces an issue that it cannot reflect local variables of the Facial-Components to the Feature-Vector. There are many components of face which contains more complex information such as eyes and mouth are more informative than forehead and cheeks.

In hybrid feature based approach, combines geometric and appearance features to overcome the issues occurred in features or appearance-based approaches. So that it can give best and efficient results [20,27]. In 3D-image representation, dynamic and static – systems are different due to the type of data, nature of data and so on. Basically, in static systems, features are extracted from many Statistical-Models. This model is Active Shape Model, Distance Based Mode, Deformable Model and so on [21]. Whereas in dynamic systems, 3D image – sequencing is utilized, also used 3D-Motion-Based features. 3-D based approaches performing well than 2D. But its quite obvious that they are facing certain issues such as High- Computation because of high resolution, frame rate and all gathered information's [22,23].

Researchers tried their experience on images which are clicked in visible – light- spectrum (VIS). [24] used Near infrared (NIR) & LBP-TOP feature description. In the component based facial features, it combines all gathers information of geometric and appearance. In general, SVM, sparse based classifier is used for recognition purpose.
[25,26] have used infrared thermal videos or images with AdaBoost algorithm and KNN classifier. [28] have recognized base on depth without using camera. This research uses local movements of face & gesture using the relations. [29-30] used sensor to detect face- region based on depth information and AAM (Active Appearance Model) to track the detected face. Basically, AAM is used for adjusting the face and Texture Model. AAM and Fuzzy logic uses prior knowledge to recognize. [31] proposed a scheme uses the colour and depth information using the Kinect sensor. This work extracted the Facial-Feature-Points Vector & recognize 6 basic emotion using a Random Forest algorithm.

In conventional approaches, features and classifiers are determined through the experts. Many approaches such as HOG, LBP, SVM, AdaBoost for same purposes. Conventional approaches generally required low power and memory than the advanced or say deep-learning approaches. These are the reason study more on these traditional methods. Also, these traditional methods give better and accurate results [32].

b. Deep Learning Based approaches

In the current scenario, there has been a breakthrough in deep learning algorithms. These algorithms are Applied in nature which include CNN, RNN for Computer Vision field, basically in deep learning-based algorithm, feature extraction, classification and last recognition. Main advantage within based approaches are that they remove completely or highly reduce the dependency on other processing techniques. This can be done through enabling end to end learning directly from the input images [33,34]. That is why in many real time issues with machine learning or computer vision face object recognition, facial expression recognition, gesture recognition, seen understanding achieve good result.

Deep learning-based application such as CNN used to boost the power of recognition there are three layers which all have a heterogeneous in nature; first is convolution layer which is used to take image as input, convert entered input with some set of filter books. these inputs inside throw the sliding window manners. After process, output features maps to represents spatial arrangements of the weights of convolutional filters within a feature map are shared. these all input feature map layers are locally connected. In the second layer known as subsampling which lower the spatial resolution of the given input feature maps to reduce their dimensions. The last layer is fully connected layer used to capture the entire original image. mostly recognition algorithm based on deep learning approach specifically CNN based [35].

Breuer and kummel use CNN based visualization techniques. they worked on some model to check various available data sets and the capability of trained network.[36,37] used two different types of CNN based approaches .in the first type, it extracts the temporal appearance features from the input image sequences .in the second type of method, it extracts temporal gesture features from the temporal facial landmark points. these two types of CNN based approach gives a better performance in recognition process.

[38] proposed approach which was a unified deep-n/named deep region and multipliable learning (DRML). this approach uses a forward function for facial regions. it behaves like a region layer, also forces the learn weights which is used to capture structural information of the image. the overall complete network is end to end trainable and automatically learned network.

This paper is already mentioned that many recognition approaches adopted deep learning models like CNN based to get better output. there are some major challenges with CNN based methods [39,40]. The major is it can reflect any temporal variations of any component. to overcome such issues a CNN based hybrid approach which is a combination of spatial features of individual frame and long –short term memory (LSTM) for the temporal features of corrective frames was developed. LSTM concept are lo ng term dependency using the short-term memory. LSTN is a special type of recurrent neural network.it is a chain like structure format for four repeating modules of a NN -

(i) CELL STATE - we can add or remove information to the cell – state.
(ii) FORGET GATE LAYER- used to *decide* which information is used to store or not.
(iii) INPUT GATE LAYER – used to decide which values should be updated or not in the cell.
(iv) OUTPUT GATE LAYER –used to provide o/p based on the current cell-state.

In compare of standalone approaches, LSTM –models are fine-tuned end to end based models. It is the basically a straight forward in nature. Also, it supports fixed as well as variable length input or output. The combined study of LSTM and CNN for recognition is more perfect and effective. Some papers proposed a hybrid RNN-CNN framework. This paper uses a continuously valued hidden layer reprobate for propagative information. This work presented a complete system which proves that a hybrid CNN-RNN architecture can perform better them the easier approaches.

There was a paper which used to tell about the spatial image characteristics of the participate expression state frames are learned using RNN similarly in the second segment of the same pear temporal characteristics of the spatial features are represented & served using LSTM chum et al proposed multilevel facial AV detection alga which is a combination of spatial & temporal features there spatial features are extracted through a CNN to reduce handcrafted description and for a temporal features LSTM are used to stack on the top of these representation doesn't matter of I/p length. Combined both o/p to get fusion.

The study suggested about 3D inception – ResNet architecture followed but LSTM unit which together extracted spatial relation as well as temporal relation, within the facial image with in the distant image in a sequence of video in this network, facial landmark point is used as input by focusing the significance facial components rather than facial region that may not contribute in generation that facial expression

[41] utilize a recurrent network to contemplate the temporal dependencies which is present in that image sequence during categorization. The experimented results used to two types of LSTM which proves that by directional networks gives a significantly better performance rather than a unidirectional LSTM.

It is suggested that a triangle optimal pattern based deep learning (MAOP-DL) method is used for solving the problem of instant change in illumination. Initially this approach omits the background and segregates the foreground from the images and then extracts the texture patterns and the relevant key attributes of the facial points.

Commonly deep learning-based approaches, establish the features and classifies by deep neutral network experts, unlike conventional methodologies. Deep Learning based approaches extract optimal attributes with the appropriate characteristics directly from the data using deep conventional neural networks. Although, Deep learning-based approaches requires a higher-level computing device and that is why it is important to minimize the computational burden at the inference time of deep learning algorithm.

As per the review conducted [17,18] the general frameworks of the hybrid CNN-LSTM and the CNN-RNN based FER approaches have the same structures. In short, the major framework of CNN-LSTM(RNN)is to amalgamate and LSTM with the deep hierarchical visual feature extractors like CNN model. Hence, this hybrid model can learn to grant and synthesis temporal dynamics for the task that involve sequential images. As described earlier, each visual feature is dissolving through a CNN is passed to LSTM and produce a fixed length vector representation. Then finally the outputs are passed to a recurrent sequence Learning model. Finally, the anticipated distribution is calculated by applying software.

7. Conclusion

This pear is presented a description of deep learning method for Facial expression and gesture recognition. In facial expression and gesture recognition, there are two types of approaches, one is known as conventional approaches and other is Based on Deep learning. In this papered explored CNN based approaches in deep learning to understand the model which learned through various datasets and demonstrated trained network based on Deep Learning Approaches. During study, it was found that hayride approaches are used to deal spatial and temporal features during recognition process so this CNN-LSTM (RNN) architecture used as hybrid approach to deal with the issues with occurred during studies.

Earlier method faces the issue of real time like they need large scale data set, massive computing power, large amount of memory and required more time for trained and test networks. So, hybrid approach is more approachable than existing. In hybrid approach which is novel in nature, is a multi-stage recurrent architecture, which involves multiple stages as its name suggested. In the first or foremost stage, the prosed CNN-LSTN (RNN) model focuses on global context aware future and on the next stage, it combines the result which occurred from the first stage with the localized and action aware. Expression and gesture estimation are computationally expensive and erroneous but overall it is fast and reliable also it is more approachable due to its code sharing nature. Analysis of review say the deep learning techniques are more successful due to its high performance. Many applicable areas of deep learning are signal processing, effective computing, HCI, Expression recognition and so on.

References

[1] X. Wang, A. Farhadi, and A. Gupta. Actions ˜ transformations. CoRR, abs/1512.00795, 2015.

[2] Y. Wang and M. Hoai. Improving human action recognition by nonaction classification. CoRR, abs/1604.06397, 2016.

[3] Z. Wang, L. Wang, W. Du, and Y. Qiao. Exploring fisher vector and deep networks for action spotting. In CVPRW, pages 10–14, 2015.

[4] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. abs/1506.01929, Dec 2015.

[5] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandr´ea, C.-E. Bichot, C. Garcia, and

[6] B. Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. CVIU, 127:14–30, Oct. 2014.

[7] D. Wu, L. Pigou, P. J. Kindermans, N. LE, L. Shao, J. Dambre, and J. M. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE TPAMI, PP (99):1–1, feb 2016.

[8] J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In ICMI, pages 453–460, 2013.

[9] J. Wu, P. Ishwar, and J. Konrad. Two-stream CNNs for gesture-based verification and identification: Learning user style. In CVPRW, 2016.

[10] X. Xu, T. M. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritized data augmentation. In Proc. ECCV, 2016.

[11] Z. Xu, L. Zhu, Y. Yang, and A. G. Hauptmann. Uts-cmu at THUMOS 2015. CVPR THUMOS Challenge, 2015, 2015.

[12] Y. Ye and Y. Tian. Embedding sequential information into spatiotemporal features for action recognition. In CVPRW, 2016.

[13] Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. CoRR, abs/1511.06984, 2015.

[14] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, et al. An introduction to computational networks and the computational network toolkit. Technical report, TR MSR, 2014.

[15] J. Yuan, B. Ni, X. Yang, and A. Kassim. Temporal action localization with pyramid of score distribution features. In CVPR, 2016.

[16] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In CVPR, pages 4694–4702, 2015.

[17] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector CNNs. CoRR, abs/1604.07669,2016.

[18] S. Zhao, Y. Liu, Y. Han, and R. Hong. Pooling the convolutional layers in deep convents for action recognition. arXiv preprint arXiv:1511.02126, 2015.

[19] T. Zhou, N. Li, X. Cheng, Q. Xu, L. Zhou, and Z. Wu. Learning semantic context feature-tree for action recognition via nearest neighbor fusion. Neurocomputing, 201:1–11, 2016.

[20] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In CVPR, pages 3323–3331, 2015.

[21] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao. A key volume mining deep framework for action recognition. In CVPR, 2016.

[22] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. reprint arXiv:1603.07772, 2016.

[23] Graves, A.; Mayer, C.; Wimmer, M.; Schmidhuber, J.; Radig, B. Facial expression recognition with recurrent neural networks. In Proceedings of the International Workshop on Cognition for Technical Systems, Santorini, Greece, 6–7 October 2008; pp. 1–6.

[24] Jain, D.K.; Zhang, Z.; Huang, K. Multi angle optimal pattern-based deep learning for automatic facial expression recognition. Pattern Recognition. Lett. 2017, 1, 1–9.

[25] Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.; Liu, Y.J.; Chen, Y.H.; Fu, X. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLoS ONE 2014, 9, e86041.

[26] Zhang, X.; Yin, L.; Cohn, J.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; Girard, J. BP4D-Spontaneous: A high resolution spontaneous 3D dynamic facial expression database. Image Vis. Comput. 2014, 32, 692–706.

[27] KDEF. Available online: http://www.emotionlab.se/resources/kdef (accessed on 27 November 2017).

[28] Die große MPI Gesichtsausdruckdatenbank. Available online: https://www.b-tu.de/en/graphic-systems/databases/the-large-mpi-facial-expression-database (accessed on 2 December 2017).

[29] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Mateo, CA, USA, 20–25 August 1995; pp. 1137–1143.

[30] Ding, X.; Chu, W.S.; Torre, F.D.; Cohn, J.F.; Wang, Q. Facial action unit event detection by cascade of tasks. In Proceedings of the IEEE International Conference Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2400–2407.

[31] Huang, M.H.; Wang, Z.W.; Ying, Z.L. A new method for facial expression recognition based on sparse representation plus LBP. In Proceedings of the International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010; pp. 1750–1754.Sensors 2018, 18, 401 20 of 20

[32] Zhen, W.; Zilu, Y. Facial expression recognition based on local phase quantization and sparse representation. In Proceedings of the IEEE International Conference on Natural Computation, Chongqing, China,29–31 May 2012; pp. 222–225.

[33] Zhang, S.; Zhao, X.; Lei, B. Robust facial expression recognition via compressive sensing. Sensors 2012,12, 3747–3761.

[34] Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 915–928.

[35] Jiang, B.; Valstar, M.F.; Pantic, M. Action unit detection using sparse appearance descriptors in space-time video volumes. In Proceedings of the IEEE International Conference and Workshops on Automatic Face & Gesture Recognition, Santa Barbara, CA, USA, 21–25 March 2011; pp. 314–321.

[36] Lee, S.H.; Baddar, W.J.; Ro, Y.M. Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos. Pattern Recognition. 2016, 54, 52–67.

[37] Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply learning deformable facial action parts model for dynamic expression analysis. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 143–157.

[38] Liu, M.; Li, S.; Shan, S.; Chen, X. Au-aware deep networks for facial expression recognition. In Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Shanghai, China, 22–26 April 2013; pp. 1–6.

[39] Liu, M.; Li, S.; Shan, S.; Chen, X. AU-inspired deep networks for facial expression feature learning. Neurocomputing 2015, 159, 126–136.

[40] Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the IEEEWinter Conference on Application of Computer Vision, Lake Placid, NY, USA, 7–9 March 2016; pp. 1–10.

[41] Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, L.; Wang, G.; et al. Recent advances in convolutional neural networks. Pattern Recognition. 2017, 1, 1–24.