Multi-sensor Data Fusion for Wearable Devices

Tiziana Rotondo
Department of Mathematics and Computer Science, University of Catania, Italy

1 RESEARCH PROBLEM

The real time information comes from multiple sources such as wearable sensors, audio signals, GPS, etc. The idea of multi-sensor data fusion is to combine the data coming from different sensors to provide more accurate information than that a single sensor alone. To contribute to ongoing research in this area, the goal of my research is to build a shared representation between data coming from different domains, such as images, signal audio, heart rate, acceleration, etc., in order to predict daily activities. In the state of the art, these arguments are treated individually. Many papers, such as (Lan et al., 2014; Ma et al., 2016) et al., predict daily activity from video or static image. Others, such as (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2014) et al., build a shared representation then rebuild the inputs or rebuild a missing modality, or (Nakamura et al., 2017) classifies from multimodal data.

2 OUTLINE OF OBJECTIVES

In the real world, the information comes from different channels, like videos, sensors, etc. Multimodal learning aims to build models that are able to process information from different modalities, semantically related, creating a shared representation to improve accuracies than could be achieved by the use a single input. As shown in Figure 1, given the images of a butterfly and a tiger and the word “butterfly ”, we want to project these data in a representation space that takes account of their correlation.

As reported in (Srivastava and Salakhutdinov, 2014), each modality is characterized by different statistical properties that don’t allow us to ignore the fact that it comes from specific input channel. The different inputs have a different representation, therefore, for a model, it is difficult to find a highly non linear relationship between different data. A good model of multimodal learning must satisfy certain properties, in fact the shared representation must be such that resemblance in the space of representation implies similarity of the corresponding inputs, to be easily obtained, even in the absence of some modalities and to fill out missing forms, starting from those observed.

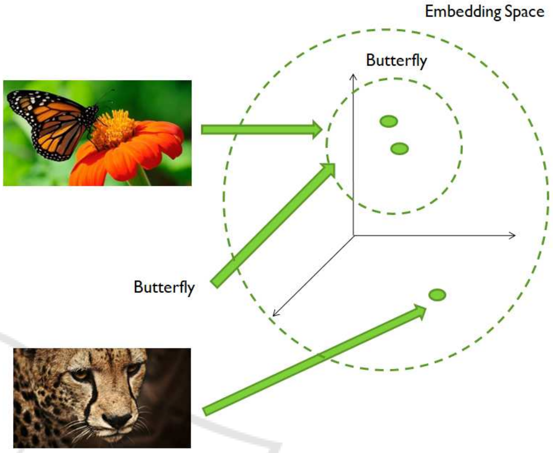


Figure 1: Idea of multimodal learning.

The problem concerning how to build a shared representation is not new (Boström et al., 2007). Fusing information is a core ability for humans. They combine all senses, sight, smell, sound, taste and touch data, for example, to understand if a food is hot or cold or in general to capture information. Sensors have been proposed also to emulate this human capability. This allows several applications in robotics, in surveillance, artificial intelligence and so on. As already mentioned, we plan to fuse multimodal learning with prediction of daily activities. The prediction of the future is a challenge that has always fascinated human people. As reported in (Lan et al., 2014), given a short video or a image, humans can predict what is going to happen in the near future. Observing the previous actions, this is possible. The creation of machines that anticipate future actions is an issue in Computer Vision field. In the state of the art, there are many applications in robotics and health care that use this predictive characteristic. For example, (Chan et al., 2017) proposed a RNN model for anticipating accidents in dashcam videos. (Koppula and Saxena, 2016) studied how to enable robots to anticipate human-object interactions from visual input in order to provide adequate assistance to the user. (Kop-

Rotondo, T.
Multi-sensor Data Fusion for Wearable Devices.
In *Doctoral Consortium (DCETE 2018)*, pages 22-28

pula et al., 2016; Mainprice and Berenson, 2013; Duarte et al., 2018) study how to anticipate human activities for improving the collaboration between human and robot.

3 STATE OF THE ART

We focus our review of related works addressing representation, frame anticipation, object interaction, action anticipation, multimodal learning, multimodal dataset and adopted system in the state of the art.

3.1 Representation

In (Vondrick et al., 2015), they explore how to anticipate human actions and objects by learning from unlabeled video. In particular, they proposed a deep networks to predict the visual representation of images in the future. In (Bütepage et al., 2017), it is proposed a deep learning framework for human motion capture data that learns a generic representation from a large corpus of motion capture data and generalizes well to new, unseen, motions, using an encoding-decoding network that learns to predict future 3D poses from the most recent past. (Ryoo et al., 2014) introduces a new feature representation named pooled time series that is based on time series pooling of feature descriptors. (Oh et al., 2015) considers spatio-temporal prediction problems where future image-frames depend on control variables or actions as well as previous frames.

3.2 Frame Anticipation

In (Vondrick and Torralba, 2017) they develop a model for generating the immediate future in unconstrained scenes that generates the future by transforming pixels in the past. In (Walker et al., 2017), the authors use the future poses generated to a Generative Adversarial Network (GAN) to predict the future frames of the video in pixel space. In (Xue et al., 2016), they propose a novel approach that models future frames from a single input image in a probabilistic manner.

3.3 Object Interaction

In (Furnari et al., 2017), it is investigated the topic of next-active-object prediction from First Person videos. They analysed the role of egocentric object motion in anticipating object interactions and propose a suitable evaluation protocol. In (Koppula and Saxena, 2016), the goal is to enable robots to predict human-object interactions from visual input in order to assist humans in daily tasks.

3.4 Action Anticipation

The goal of action anticipation is to detect an action before it happens. In (Gao et al., 2017), it is proposed a Reinforced Encoder-Decoder (RED) network for action anticipation that takes multiple history representations as input and learns to anticipate a sequence of future representations. These anticipated representations are processed by a classification network for action classification. In (Lan et al., 2014), it is presented a hierarchical model that represent the human movements to infer future actions from a static image or a short video clip. In (Ma et al., 2016), the authors proposed a method to improve training of temporal deep models to learn activity progression for activity detection and early recognition tasks.

3.5 Multimodal Learning

The aforementioned papers mostly concern a single modality such as video. We want to extend these concepts to multimodal inputs. This problem is not only theoretical but it has already been dealt with machine learning techniques. One of the firsts paper on Multimodal Learning is (Ngiam et al., 2011) where video and audio signals are used as input. The aims of this article are: compress the inputs into a shared representation and then rebuild them and rebuild a missing mode, for example from the video, you want to get the audio signal and the video signal as output. The creation of a fused representation has also been treated in other papers (Srivastava and Salakhutdinov, 2014; Aytar et al., 2017), in particular they build a representations that are robust in another way. Indeed, these representations are very important because they are fundamental components to understand relationships between modalities.

In (Nakamura et al., 2017), a model for reasoning on multimodal data to jointly predict activities and energy expenditures is proposed. In particular, for these tasks they consider Egocentric videos augmented with heart rate and acceleration signals. In (Wu et al., 2017), it is proposed a on-wrist motion triggered sensing system for anticipating daily intention. They introduces a Recurrent Neural Network (RNN) to anticipate intention and a policy network to reduce computation requirement.

3.6 Multimodal Dataset

In this section, we discuss about the datasets in the state-of-the-art. The data are an important point; in fact these are collected at different sampling frequencies, therefore, before proceeding to extract the features, it is necessary to synchronize the various inputs in order to have all the related modalities.

The egocentric multimodal dataset (Stanford-ECM) (Nakamura et al., 2017) comprises 31 hours of egocentric video (113 videos) augmented with acceleration and heart rate data. The video and triaxial acceleration were capture with mobile phone with a 720×1280 resolution and 30 fps and 30Hz, respectively. The lengths of the individual videos covered a diverse range from 3 minutes to about 51 minutes in length. The heart rate was collected with wrist sensor every 5 seconds (0.2 Hz). These data was time-synchronized through Bluethoot.

The Multimodal Egocentric Activity dataset (Song et al., 2016) contains 20 distinct life-logging activities performed by different human subjects and comprises these data: video, accelerometer, gravity, gyroscope, linear acceleration, magnetic field and rotation vector. The Google Glass enables to synchronize egocentric video and sensor data. The video was collected with a 1280×720 resolution and 29.9 fps while the sensor data with a lenght of 15 second and 10Hz. Each activity category has 10 sequences of 15 seconds.

Multimodal User-Generated Videos Dataset (Bano et al., 2015) contains 24 user-generated videos (70 mins) captured using hand-held mobile phones in high brightness and low brightness scenarios (e.g. day and night-time). The video (audio and visual) along with the inertial sensor (accelerometer, gyroscope, magnetometer) data is provided for each video. These recordings are captured using single camera at distinct timings and locations, changing lights and varying camera motions. Each captured video was manually annotated to get labels for camera motions (pan,tilt, shake) at each second. The ground-truth labels are included in the dataset.

In (Wu et al., 2017), the authors collected Daily Intention Dataset that was used for training model to predict the future and they select 34 daily intentions. Each of this is associated with a motion and an object. The video was collected with a 640×480 resolution.

3.7 Adopted System

In this section, we describe some of the models that are presented in the state of the art. Many of these are based on the study of different deeps networks, starting from the Restricted Boltzmann Machines (RBM) (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2014) to more used Convolutional Neural Networks (CNN) (Nakamura et al., 2017), accumulated by the fact that each architecture processes a probability distribution on all of it multimodal input space.

3.7.1 Boltzmann Machines

The Boltzmann Machines (BM) (Salakhutdinov and Hinton, 2009) are networks with a symmetrical connections between binary units, called visible variables $\mathbf{v} \in \{0,1\}^D$ and hidden variables $\mathbf{h} \in \{0,1\}^P$. There are connections between the visible state and the hidden state and between the units of the same type. The energy of the state $\{v,h\}$ is defined as

$$E(v,h;\theta) = -\frac{1}{2}v^T Lv - \frac{1}{2}h^T Jh - v^T Wh, \qquad (1)$$

where $\theta = \{W,L,J\}$ are the parameters of the model that represent, respectively, the interactions between the visible-hidden, visible-visible and hidden-hidden states. The probability that the model assigns to the visible variable $\mathbf{v}$ is

$$p(\mathbf{v};\theta) = \frac{p^*(\mathbf{v},\theta)}{Z(\theta)} = \frac{1}{Z(\theta)}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h};\theta)), \quad (2)$$

$$Z(\theta) = \sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h};\theta)). \qquad (3)$$

where p^* is the non-normalized probability and $Z(\theta)$ the partition function. Updating the parameters necessary to calculate the log-likelihood with the gradient descent method are obtained from (2):

$$\begin{aligned} \triangle W &= \alpha(E_{Pdata}[vh^T] - E_{Pmodel}[vh^T]),\\ \triangle L &= \alpha(E_{Pdata}[vv^T] - E_{Pmodel}[vv^T]),\\ \triangle J &= \alpha(E_{Pdata}[hh^T] - E_{Pmodel}[hh^T]), \end{aligned} \qquad (4)$$

where α is the learning rate, $E_{Pdata}[\cdot]$ is the data dependency prediction and $E_{Pmodel}[\cdot]$ is the prediction on the model. The learning algorithm of the BMs requires a very long execution time because it is necessary to initialize in a random way the Markov chains to estimate the predictions on the data and on the model. Learning is more effective if you use the Restricted Boltzmann Machines (Srivastava and Salakhutdinov, 2014; Salakhutdinov and Hinton, 2009) (RBMs). In such models there are connections between the visible layer and the hidden state but there are no connections between variables of the same type. The parameters L,J are null. In this case the algorithm is efficient using the Contrastive Divergence which provides an approximation of the log-likelihood with a short Markov chain. It is possible to use a stochastic approximation to approximate the prediction of the model. θ_t and X_t, respectively, the parameter and the status are added as follows:

- Given X_t, X_{t+1} is updated by an operator $T_\theta(X_{t+1}, X_t)$ leaving p_{θ_t} unchanged.
- θ_{t+1} is obtained by replacing the predictability of the intractable model with the prediction against X_{t+1}.

A necessary condition for convergence is that the learning rate decreases as time passes $\sum_{t=0}^{max} \alpha_t = +\infty$ and $\sum_{t=0}^{max} \alpha_t^2 < +\infty$. This is satisfied for $\alpha_t = 1/t$.

The models described are the base cell of the Deep Boltzmann Machines (DBM) (Salakhutdinov and Hinton, 2009). These latest networks allow us to learn the potential of internal representations and allow us to deal with unlabelled or partially labelled data.

In (Ngiam et al., 2011), RBMs are used to build a shared representation, as shown in Figure 2. One of the most linear RBM approaches for audio and video, as in Figure 2.a, 2.b. The resulting probability can be used as a new representation of the data. This method is used as a reference model.

The 2.c model was given input to the concatenation of inputs, but given the nonlinear correlation of the data, it is difficult for the RBM to provide a multimodal representation. In particular, the units born have strong connections between the individual modes and weak connections between units that connect the two modes.

Finally, a model that takes into account the previous ones is considered; in fact, the modalities are trained separately and then the results are concatenated. The first level of visualization is a phoneme and at my levels. The latter model is used to train weights to use autoencoder models.

3.7.2 RNN

The RNNs (LeCun et al., 2015) are models that process an input sequence one element at a time, keeping in its hidden units a state vector that contains information related to the previous elements of the sequence. These networks use the same parameters, for this reason they perform the same computation at each moment of time on different inputs of the same sequence. The training of such networks is affected by the vanishing/exploding gradient problem, in which the calculated and propagated backward gradients tend to increase or decrease at each moment of time, therefore, after a certain number of instants of time, the gradient diverges to infinity or converges to zero. To overcome this problem, long short-term memory (LSTM) networks are used, which are particular RNNs with hidden units that recall previous inputs for a long time. Such units, taking as input, at each instant, the previous state and the current one and combining them, decide which information to keep and which to delete from memory.

The aim of the paper (Nakamura et al., 2017) is to recognize a daily activity and calculate its energy expenditure, starting from a multimodal dataset. An LSTM is introduced that takes in input a multimodal representation of the video and the acceleration and returns in output the activity label and the energy consumption for each frames. Heart rate is also integrated to estimate the energy spent.

In (Wu et al., 2017), a model based on the RNN is proposed with two LSTM layers in order to be able to handle the variations, as follows

$$g_t = Emb(W_{emb}, con(f_{m,t}, f_{o,t})), \tag{5}$$

$$h_t = RNN(g_t, h_{t-1}), \tag{6}$$

$$p_t = Softmax(W_y, h_t), \tag{7}$$

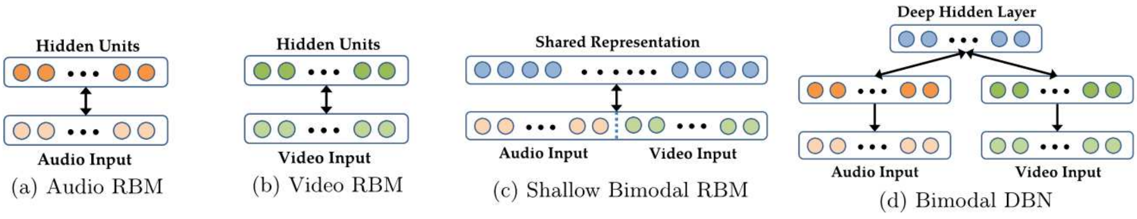
$$y_t = arg \max_{y \in Y} p_t(y), \tag{8}$$

where Y is the set of intent indices, p_t is the softmax probability of each intention in Y, W_y is the parameter of the model to train, h_t is the hidden representation coached, g_t is the fixed size of the output of $Emb(\cdot)$, W_{emb} is the parameter of the embedding function $Emb(\cdot)$, $con(\cdot)$ is the concatenation operation and $Emb(\cdot)$ is a linear mapping function.

Policy network π is also introduced to determine when to process an image in a representation of the f_o object. The network continuously observes the movement $f_{m,t}$ and the hidden state of the RNN h_t to be able to calculate $f_{o,t+1}$.

4 EXPECTED OUTCOME

In this section, we describe our pipeline, shown in Figure 3. Stanford-ECM Dataset (Nakamura et al., 2017) is considered. It has video, acceleration and heart rate data, so the problem is defined as follows: given $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T\}$ with $\mathbf{v}_i \in \mathbb{R}^2$ $\forall i \in \{1, 2, ..., T\}$ a sequence of video frame, $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_T\}$ with $\mathbf{a}_i \in \mathbb{R}^3$ $\forall i \in \{1, 2, ..., T\}$ a sequence of acceleration signals and $\mathcal{HR} = \{hr_1, hr_2, ..., hr_T\}$ with $hr_i \in \mathbb{R}$ $\forall i \in \{1, 2, ..., T\}$ a sequence of heart rate signal, (x_t^v, x_t^a, x_t^{hr}) $\forall t \in \{1, 2, ..., T\}$ are the feature representations of video, acceleration and heart rate signal and $x_t = (x_t^v, x_t^a, x_t^{hr})^T$ the features vector. Given $(x_t, x_{t+1}, label_t, label_{t+1})$ as input, we want to predict next action, observing only data before the activity starts.



(a) Audio RBM (b) Video RBM (c) Shallow Bimodal RBM (d) Bimodal DBN

Figure 2: RBM Pretraining Models.

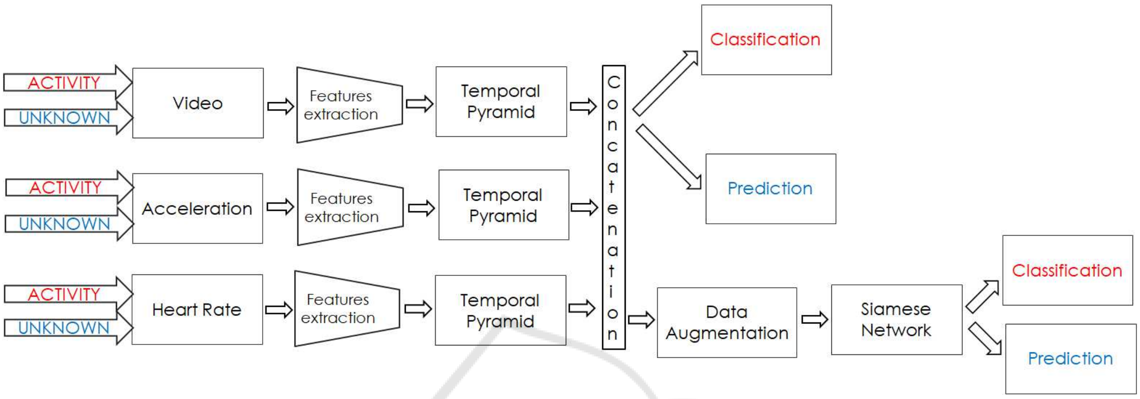


Figure 3: Pipeline of our model.

Since this dataset was created for classification task, it is not specific for prediction task. It is adapted for our task and Unknown/Activity transitions are selected. The dataset is cut around transitions and 64 frames before and after transitions are considered. Now, we describe how feature representation x_t^v, x_t^a and x_t^{hr} for each signal has been obtained.

For visual data, features are extracted from pool5 layer of Inception CNN, pretrained on ImageNet (Deng et al., 2009). Each video frame has been transformed into a 1024-dimensional feature vector x_t^v.

For acceleration data, the features have been extracted from the raw signal and a sliding window with size of 32fps has been considered. For time-domain features, mean, standard deviation, skewness, kurtosis, percentiles (10th, 25th, 50th, 75th, 90th), acceleration count for each axis and correlation coefficients between each axis have been computed. For frequency-domain features, we consider the spectral entropy $J = -\sum_{i=0}^{N/2} \bar{P}_i \cdot \log_2 \bar{P}_i$ where $\bar{P}_i$ is the normalized power spectral density computed from Short Time Fourier Transform (STFT). Then, all features are concatenated and x_t^a is a 36-dimensional vector.

For hear rate data, the features are extracted from the time-series of the raw signals. Mean and standard deviation are computed and $x_t^{hr} \in \mathbb{R}^2$.

Features are represented in a temporal pyramid (Pirsiavash and Ramanan, 2012) with three levels (level 0, level 1 and level 2). The top level j = 0 is a histogram (mean) over the full temporal extent of a data, the next level (j=1) is the concatenation of two histograms obtained by temporally segmenting each modality into two halfs, and so on. In this way, we have 7 histograms. Therefore, 1024 × 7 visual features, 36 × 7 acceleration features, and 2 × 7 heart rate features are obtained. All features are concatenated into a single vector $x_t = (x_t^v, x_t^a, x_t^{hr})^T$ and features vector size is 7434.

Since we have a few data, a data augmentation technique is used to expand the training set to prevent over-fitting. As reported in (Krizhevsky et al., 2012) geometric transformation and RGB channels alteration are the traditional data augmentation approaches, while (Zhu et al., 2017) proposes a method with GAN and CycleGAN.

In the current configuration, the permutation is considered of each piece of unknown with the various sections of the activities. For Siamese Network, each unknown is paired with any activity. In this case, we created a "false" couples but these are necessary to implement the Siamese network. The labels of each transition are changed from 0-8 to 0-1, as follows: if unknown and the activity belong to the same class (e.g. unknown related to walking and walking), 1 is assigned, otherwise 0 is assigned if unknown and the activity are different (e.g. unknown related to walking and food preparation).

The obtained dataset is strongly unbalanced, so we considered 121 sequences from same unknown and activity and 154 sequences from different unknown and activity for each class, in order to balance activity classes and unknown class. In this way, dataset has 12177 sequences. This is the input for Siamese network (Koch et al., 2015) that consists of twin networks which accept distinct inputs but the weights are shared. During training the two sub-networks extract features from two inputs, while the joining neuron measures the distance between the two feature vectors.

In our experiment, euclidean metrics is used to calculate the distances between inputs. The contrastive loss function has been used. Three convolutional layers are considered with numbers of filters 32, 64 and 2, respectively, all of size 3×1 and a relu activation function. The output of each convolutional layer is reduced in size using a max-pooling layer that halves the number of features. A k-nearest-neighbor classification algorithm (K-NN) and a support vector machine (SVM) are used on features for classification purposes.

5 STAGE OF THE RESEARCH

Our preliminary results prove that we can predict daily activity from multimodal data. In particular, Stanford-ECM Dataset has been considered and we implemented a siamese network to build an embedding space. The performance of the experiment has been evaluated with a SVM for different kernels and a K-NN for different values of K.

In future works, we want to improve our pipeline and test its on other datasets. The result that I expect and that should validate the problem and the approach is to overcome the values of accuracy in classification baseline.

ACKNOWLEDGEMENTS

I would like to thank my advisors Prof. Sebastiano Battiato (University of Catania), Prof. Giovanni Maria Farinella (University of Catania) and Ing. Valeria Tomaselli (STMicroelectronics, Catania) for their continued supervision and support.

REFERENCES

Aytar, Y., Vondrick, C., and Torralba, A. (2017). See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932.

Bano, S., Cavallaro, A., and Parra, X. (2015). Gyro-based camera-motion detection in user-generated videos. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1303–1306, New York, NY, USA. ACM.

Boström, H., Andler, S. F., Brohede, M., Laere, J. V., Niklasson, L., Nilsson, M., Persson, A., and Ziemke, T. (2007). On the definition of information fusion as a field of research. Technical report.

Bütepage, J., Black, M. J., Kragic, D., and Kjellström, H. (2017). Deep representation learning for human motion prediction and classification. *CoRR*, abs/1702.07486.

Chan, F.-H., Chen, Y.-T., Xiang, Y., and Sun, M. (2017). Anticipating accidents in dashcam videos. In Lai, S.-H., Lepetit, V., Nishino, K., and Sato, Y., editors, *Computer Vision – ACCV 2016*, pages 136–153, Cham. Springer International Publishing.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Duarte, N., Tasevski, J., Coco, M. I., Rakovic, M., and Santos-Victor, J. (2018). Action anticipation: Reading the intentions of humans and robots. *CoRR*, abs/1802.02788.

Furnari, A., Battiato, S., Grauman, K., and Farinella, G. M. (2017). Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401 – 411.

Gao, J., Yang, Z., and Nevatia, R. (2017). RED: reinforced encoder-decoder networks for action anticipation. *CoRR*, abs/1707.04818.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition.

Koppula, H. S., Jain, A., and Saxena, A. (2016). *Anticipatory Planning for Human-Robot Teams*, pages 453–470. Springer International Publishing, Cham.

Koppula, H. S. and Saxena, A. (2016). Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):14–29.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lan, T., Chen, T.-C., and Savarese, S. (2014). A hierarchical representation for future action prediction. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 689–704, Cham. Springer International Publishing.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deel learning. *Nature*, 521(7553):436–444.

Ma, S., Sigal, L., and Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1942–1950.

Mainprice, J. and Berenson, D. (2013). Human-robot collaborative manipulation planning using early prediction of human motion. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 299–306.

Nakamura, K., Yeung, S., Alahi, A., and Fei-Fei, L. (2017). Jointly learning energy expenditures and activities using egocentric multimodal signals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6817–6826.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 689–696, USA. Omnipress.

Oh, J., Guo, X., Lee, H., Lewis, R. L., and Singh, S. P. (2015). Action-conditional video prediction using deep networks in atari games. *CoRR*, abs/1507.08750.

Pirsiavash, H. and Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854.

Ryoo, M. S., Rothrock, B., and Matthies, L. H. (2014). Pooled motion features for first-person videos. *CoRR*, abs/1412.6505.

Salakhutdinov, R. and Hinton, G. E. (2009). Deep boltzmann machines. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida*, pages 448–455.

Song, S., Cheung, N., Chandrasekhar, V., Mandal, B., and Lin, J. (2016). Egocentric activity recognition with multimodal fisher vector. *CoRR*, abs/1601.06603.

Srivastava, N. and Salakhutdinov, R. (2014). Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980.

Vondrick, C., Pirsiavash, H., and Torralba, A. (2015). Anticipating the future by watching unlabeled video. *CoRR*, abs/1504.08023.

Vondrick, C. and Torralba, A. (2017). Generating the future with adversarial transformers. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2992–3000.

Walker, J., Marino, K., Gupta, A., and Hebert, M. (2017). The pose knows: Video forecasting by generating pose futures. *CoRR*, abs/1705.00053.

Wu, T., Chien, T., Chan, C., Hu, C., and Sun, M. (2017). Anticipating daily intention using on-wrist motion triggered sensing. *CoRR*, abs/1710.07477.

Xue, T., Wu, J., Bouman, K. L., and Freeman, W. T. (2016). Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. *CoRR*, abs/1607.02586.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593.