The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Recognizing Human Interactions Using Group Feature Relevance in Multinomial Kernel Logistic Regression

Ouiza Ouyed, Mohand Said Allili

Université du Québec en Outaouais
Département d'informatique et d'ingénierie.
Gatineau, J8X 3X7, QC, Canada.
Emails: {ouiza.ouyed, mohandsaid.allili}@uqo.ca

Abstract

We propose a supervised approach incorporating group feature sparsity in multi-class kernel logistic regression (GFR-MKLR). The need for group sparsity arises in several practical situations where a subset of a set of factors can explain a predicted variable and each factor consists of a group of variables. We apply our approach for predicting human interactions based on body parts motion (e.g., hands, legs, head, etc.) where image features are organised in groups corresponding to body parts. Our approach, leads to sparse models by assigning weights to groups of features having the highest discrimination between different types of interactions. Experiments conducted on the UT-Interaction dataset have demonstrated the performance of our method with regard to stat-of-art methods.

Introduction

Models with sparsity constraint on solutions plays a central role in many high-dimensional classification problems (Hastie et al. 2015; Tibshirani 1996). In some cases, explanatory variables can be grouped together into separate factors influencing prediction of classes (Rao et al. 2016). This is the case, for example, in real world human activities captured in videos where a single activity can be decomposed into co-occurring actions performed by different persons (e.g., hand shaking, hugging, meeting, etc.) (Noceti et al. 2014). Each action, performed by the person, incurs the motion of different body parts depending on the gestures performed in the action (Aggarwal et al. 2011). Therefore, having sparse classification models selecting features at the gesture level is an important issue for better activity recognition. Sparse models usually prevent over-fitting and lead to more interpretable solutions in high-dimensional machine learning problems (Hastie et al. 2015; Rao et al. 2016).

Group sparsity has been proposed in the past mainly as an extension to the *least absolute shrinkage and selection operator* (LASSO) method (Tibshirani et al. 1996; Yuan et al. 2006). Contrarily to LASSO which performs feature selection for individual features, group LASSO performs selection for entire groups of variables, where each group constitutes a separate explanatory factor (Hastie et al. 2015). In particular, it can be assumed that the optimal sparsity will tend to involve clusters or groups of coefficients, corresponding to preexisting groups of features (Yuan et al. 2006). While the form of the groups can be a priori known (e.g., in activity recognition, a group can correspond to all features associated with a part of the body performing a gesture), the subset of groups that is relevant to the classification task at hand can be unknown. Recently, group LASSO methods have enjoyed a tremendous success in high dimensional classification problems (Vincent et al. 2014; Wang et al. 2008; Wang et al. 2013). It remains, however, that most proposed methods are limited to binary classification based on linear models.

One of the earliest work on two person interaction recognition using motion trajectories was proposed by Datta *et al.* (Datta et al. 2002). This method tracks the trajectories of different parts of the body, then tries to dissociate violent from non-violent actions. In (Park et la. 2003), a hierarchical Bayesian network (BN) is proposed for interaction recognition. In this method, low-level nodes of the network are used to represent the pose of body parts, whereas high-level nodes estimate the overall body pose. (Ryoo et al. 2009) designed a method to measure structural similarity between sets of spatiotemporal features extracted from two videos. The authors then derived a kernel for action classification based on support vector machines (SVM). In the same vein, (Slimani et al. 2014) designed a correlation matrix between spatiotemporal features used to represent and classify interactions between persons using SVM. In (Meng et al. 2012), location and appearance of human joints are combined for interaction representation, which are then classified using SVM.

Similarly to our approach, (Yuan et al. 2012) groups trajectories of densely-sampled key points in videos to form interaction components. Then, different interactions are compared using a spatiotemporal context kernel plugged in an SVM classifier. In (Do and Pustejovsky 2017), authors propose a compact representation for human-object interactions by comparing quantitative and qualitative features at two levels: frame level features (by visualizing human and objet trajectories) and event level (by summarizing the change between first, middle and last frames across event duration). These methods, however, incorporate sparsity at the level of interaction representation, whereas our method focuses on integrating sparsity in the classification stage of interactions.

In this paper, we are interested in extending group sparsity in a multi-class setting for recognizing activities involving interactions between two individuals. Building on the success of kernel-based classification methods applied to single action recognition in videos (e.g., walking, sitting, etc.) (e.g., walking, sitting, etc.) (Aggarwal et al.2011; Schuldt et al. 2004), we propose a sparse model based on multinomial kernel logistic regression for recognition of activities involving interactions between persons. We represent motion of each moving person by tracking trajectories of key joints over the video frames. A group of features is defined for each trajectory and the concatenation of the group of features gives the final representation of each interaction. The direction of the trajectories generated by the persons are different among different types of interactions. To emphasize these differences and select the most discriminative trajectories, group of features weighting is integrated in kernel logistic regression instead of weighting each feature as in the case of LASSO. We show that our algorithm yields better results in comparison with several recent methods.

The rest of the paper is organized as follows. Section 2 provides details about interaction representation. Section 3 presents the proposed model for interaction classification. Section 4 presents some experimental results. We conclude the paper with a conclusion and future work perspectives.

Interaction representation and classification

In our method, we represent interactions using features extracted from the motion of human joints. On the constructed feature space, interaction recognition is performed using sparse multinomial kernel logistic regression. In the following section, we first describe human interaction representation. Then, we give the details of the proposed classification model based on group sparsity. The outline of the steps of our method is shown in Figure 1.

Interaction representation

For an input video with T frames, we extract the trajectory for all key points corresponding to a set of human joints $\mathbf{J}$. For this purpose, we track each joint over video frames $\mathbf{F}_t$, $t \in \{1, ..., T\}$ using the algorithm proposed in (Yang et al. 2011). There are a total of 7 joints in the following order: *head (H), right shoulder (RS), left shoulder (LS), right hand (RH), left hand (LH), right foot (RF) and left foot (LF)*, which are tracked over the video frames. The concatenation of joint locations $(l_1, l_2, ..., l_T)$ with $l_t = (x_t, y_t)$ form the interaction trajectories $\mathbf{tr}_{J_i}$, $i \in \{1, ..., 14\}$ (7 trajectories per person). In order to eliminate false joints detections over frames, we use a median filter to smooth the resulting trajectories. The static joints which give points or small trajectories are retained since the goal is to prove the implication or not of a joint movement to discriminate between different interactions. Figure 2 shows examples of extracted trajectories for *punch*, *kick* and *point* interactions. From each trajectory $\mathbf{tr}_{J_i}$, two features are computed to describe joints shape and motion. Given a trajectory of length $\mathbf{T}$, its shape is described by : $(\Delta l_1, \Delta l_2, ..., \Delta l_{T-1})$, with $\Delta l_t = (\Delta l_{x_t}, \Delta l_{y_t}) = (x_{t+1} - x_t, y_{t+1} - y_t)$. The final displacement vectors according to the coordinates x and y are normalized as follow:

$$D_{x,y} = \frac{(\Delta l_1, \Delta l_2, ..., \Delta l_{T-1})}{\sum_{j=1}^{T-1} \|\Delta l_j\|}, \quad (1)$$

The normalized histogram of displacement HOD is then obtained by concatenating the histograms of D_x and D_y as follow $HOD = [HOD_x, HOD_y]$.The motion of each trajectory is described by a local curvature in space and time coordinates, respectively, $\mathbf{x}, \mathbf{y}$ **and** $\mathbf{t}$ (Rao et al. 2002). The curvature $\mathbf{C_t}$ at each frame t is defined in Eq. (2)

$$\mathbf{C_t} = \frac{x_t^{'} y_t^{''} - y_t^{'} x_t^{''}}{(x_t^{'2} + y_t^{'2} + 1)^{3/2}}, \quad (2)$$

with, $x_t^{'}, y_t^{'}, x_t^{''} and y_t^{''}$ are the first and second order temporal derivatives of the trajectory position, with: $x_t^{'} = \Delta l_{x_t}, y_t^{'} = \Delta l_{y_t}, x_t^{''} = \Delta x_t^{'}, y_t^{''} = \Delta y_t^{'}$ *and* $\Delta t = 1$ knowing that the trajectories are extracted over successive frames. The shape and the motion of the given trajectory is then described by a concatenation of normalized histogram of displacement and curvature to form a group of features: $[HOD, HOC]$. Finally, for each video two descriptors per trajectory are concatenated following a certain spatial order starting from right to left and from up to bottom as follow: $[(HOD_{tr_H}, HOC_{tr_H})(p1), (HOD_{tr_{RS}}, HOC_{tr_{RS}})(p1), ..., (HOD_{tr_{LF}}, HOC_{tr_{LF}})(p1), HOD_{(tr_H}, HOC_{tr_H})(p2), ..., (HOD_{tr_{LF}}, HOC_{tr_{LF}})(p2)]$, p1 and p2 refers to person 1 and person 2 in a current video frame.

Interaction classification

Since each interaction is represented by a group of features of each trajectory, we aim in this section to discriminate between activities by weighting a discriminant group of features according to their contribution. This work is an extension of our previous work (Ouyed et Allili. 2014), instead of weighting individual features in a multinomial kernel logistic regression the weights are attributed for a group of features which refers to a trajectory descriptors. details of our classification method are given in section.

Group feature weighting for MKLR

Assume that we have n instances of training data $\mathbf{x}_i \in \mathbb{R}^d$, $i \in \{1, ..., n\}$, with d measured features for each instance. The features are partitioned in G groups $\mathbf{x}_i^{(g)} \in \mathbb{R}^{d_g}$, $g \in \{1, ..., G\}$ and $d_g = d/G$ and we can rewrite the full vector $\mathbf{x}_i$ as: $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, ..., \mathbf{x}_i^{(G)})$. Therefore, instead of calculating relevance at the level of each feature, we calculate relevance for each group of features g in the multinomial kernel logistic regression. Suppose that the data are generated from m classes ($m \geq 2$). We associate an encoding vector $\mathbf{y}_i = [y_i^{(1)}, y_i^{(2)}, ..., y_i^{(m)}]^T$ for each data point $\mathbf{x}_i$.

Using a weighting vector $\Psi^{(j)} = [\psi_1^{(j)}, \psi_2^{(j)}, ..., \psi_G^{(j)}]^T$ for each class $j \in \{1, ..., m-1\}$, we associate a separate symmetric kernel $\tilde{\mathbf{K}}^{(j)}$ for each class j encoding the class group of features relevance. The kernel entries for a class j are calculated as follows:

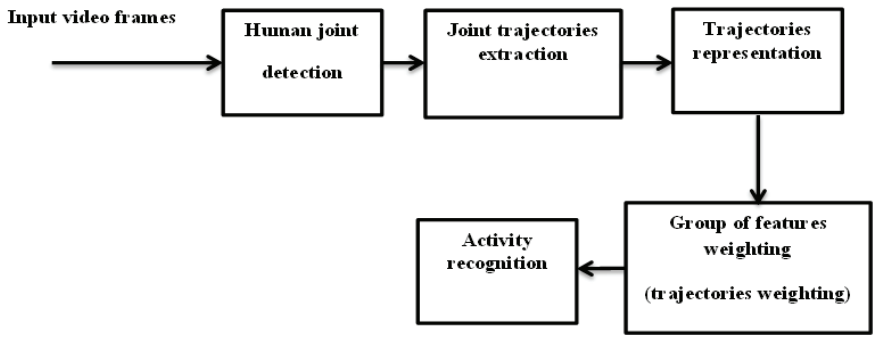
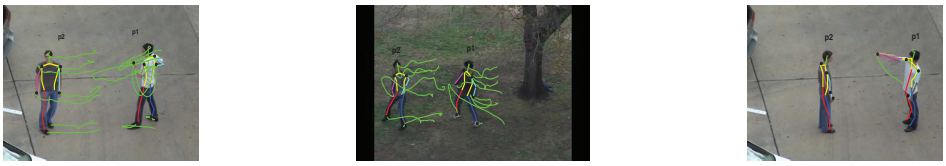


Figure 1: Our framework of human interaction recognition.



Punch Kick Point

Figure 2: Examples of extracted trajectories.

$$\tilde{\mathcal{K}}^{(j)}(\mathbf{x}_r, \mathbf{x}_s) = \exp\left(-\frac{1}{2}\sum_{g=1}^{G}\psi_g^{(j)}\|\mathbf{x}_r^{(g)} - \mathbf{x}_s^{(g)}\|\right) \quad (3)$$

Using the ℓ_0-"norm" penalization, the NLL is given as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{A}, \mathbf{\Psi}) &= \sum_{j=1}^{m-1} -\mathbf{y}^{(j)^T}\tilde{\mathbf{K}}^{(j)}\mathbf{a}^{(j)} \\ &+ \mathbf{1}^T \ln\left[1 + \sum_{h=1}^{m-1}\exp(\tilde{\mathbf{K}}^{(h)}\mathbf{a}^{(h)})\right] \\ &+ \sum_{j=1}^{m-1}\left[\frac{\lambda}{2}\mathbf{a}^{(j)^T}\tilde{\mathbf{K}}^{(j)}\mathbf{a}^{(j)}\right. \\ &+ \left.\mu\sum_{g=1}^{G}\left[1 - \exp(-\beta\psi_g^{(j)})\right]\right], \quad (4)\end{aligned}$$

where $\mathbf{A} = [\mathbf{a}^{(1)}, ..., \mathbf{a}^{(m-1)}]$ and $\mathbf{\Psi} = [\Psi^{(1)}, ..., \Psi^{(m-1)}]$. We have $\forall j \in \{1, ..., m-1\}, \forall g \in \{1, ..., G\}$:

$$\begin{aligned}\partial\mathcal{L}/\partial\mathbf{a}^{(j)} &= \tilde{\mathbf{K}}^{(j)}\mathbf{c}^{(j)} \quad (5)\\ \partial\mathcal{L}/\partial\Psi^{(j)} &= [\mathbf{c}^{(j)^T}\mathbf{Q}_1^{(j)}\mathbf{a}^{(j)}, ..., \mathbf{c}^{(j)^T}\mathbf{Q}_G^{(j)}\mathbf{a}^{(j)}]^T \\ &+ \mu\beta\exp(-\beta\Psi^{(j)}), \quad (6)\end{aligned}$$

where we define $\mathbf{c}^{(j)} = \left(-\mathbf{y}^{(j)} + \mathbf{p}^{(j)} + \frac{\lambda}{2}\mathbf{a}^{(j)}\right)$ and $\mathbf{Q}_g^{(j)} = \tilde{\mathbf{K}}^{(j)} \circ \mathbf{B}_g^{(j)}$, with $\mathbf{B}_g^{(j)}$ is an $n \times n$ matrix having entries defined by $\mathbf{B}_g^{(j)}(r, s) = -\psi_g^{(j)}\|\mathbf{x}_r^{(g)} - \mathbf{x}_s^{(g)}\|$ and $\mathbf{p}^{(j)}$ is the class posterior probability defined for each instance. The symbol $\circ$ defines the element-wise Hadamard product between matrices. It follows that the gradient of $\mathcal{L}$ with respect to the vectors $\mathbf{a}^{(j)}$'s and $\Psi^{(j)}$'s will be given by:

$$\tilde{\mathbf{g}} = \left(\tilde{\mathbf{K}}^*(\tilde{\mathbf{p}} - \tilde{\mathbf{y}} + \lambda\tilde{\mathbf{a}}), \mathbf{h}^{(1)}, \cdots, \mathbf{h}^{(m-1)}\right), \quad (7)$$

where $\mathbf{h}^{(j)} = \left[\mathbf{c}^{(j)^T}\mathbf{Q}_1^{(j)}\mathbf{a}^{(j)}, ..., \mathbf{c}^{(j)^T}\mathbf{Q}_G^{(j)}\mathbf{a}^{(j)}\right]^T + \mu\beta\exp(-\beta\Psi^{(j)})$, $\tilde{\mathbf{a}} = [\mathbf{a}^{(1)^T}, \mathbf{a}^{(2)^T}, ..., \mathbf{a}^{(m-1)^T}]^T$ and $\tilde{\mathbf{K}}^* = \text{diag}[\tilde{\mathbf{K}}^{(1)}, ..., \tilde{\mathbf{K}}^{(m-1)}]$. The operator diag$[\cdot]$ builds a matrix with diagonal blocks made of the elements of the arguments. To calculate the Hessian of function (4), note that the Hessian with respect to the elements of $\mathcal{A}$ is given by the matrix $\tilde{\mathbf{K}}^*\mathbf{W}^*\tilde{\mathbf{K}}^* + \lambda\tilde{\mathbf{K}}^*$, where we define $\tilde{\mathbf{K}}^* = \text{diag}[\tilde{\mathbf{K}}^{(1)}, ..., \tilde{\mathbf{K}}^{(m-1)}]$. We define also the matrix $\mathbf{W}^*$ as follows:

$$\mathbf{W}^* = \begin{pmatrix} \mathbf{W}_{1,1} & \mathbf{W}_{1,2} & \cdots & \mathbf{W}_{1,m-1} \\ \mathbf{W}_{2,1} & \mathbf{W}_{2,2} & \cdots & \mathbf{W}_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{m-1,1} & \mathbf{W}_{m-1,2} & \cdots & \mathbf{W}_{m-1,m-1} \end{pmatrix}, \quad (8)$$

with:

$$\mathbf{w}_{j,\ell} = \begin{cases} \text{diag}[p_1^{(j)}(1 - p_1^{(j)}), ..., p_n^{(j)}(1 - p_n^{(j)})] & \text{if} \quad j = \ell. \\ \text{diag}[-p_1^{(j)}p_1^{(\ell)}, ..., -p_n^{(j)}p_n^{(\ell)}] & \text{if} \quad j \neq \ell. \end{cases} \quad (9)$$

Similarly to the case of binary clarification, we also need to calculate matrices $\mathbf{T}^{(j)}$ and $\mathbf{M}^{(j)}$ for each class j, $j \in \{1, ..., m-1\}$, with elements defined as follows: $\mathbf{T}^{(j)} = \frac{\partial^2\mathcal{L}}{\partial\Psi^{(j)}\partial\Psi^{(j)^T}}$ and $\mathbf{M}^{(j)} = \frac{\partial^2\mathcal{L}}{\partial\mathbf{a}^{(j)}\partial\Psi^{(j)^T}}$. The full Hessian matrix with respect to all the parameters is given as follows:

$$\tilde{\mathbf{H}} = \begin{pmatrix} \tilde{\mathbf{K}}^*\mathbf{W}^*\tilde{\mathbf{K}}^* + \lambda\tilde{\mathbf{K}}^* & \mathbf{M}^* \\ \mathbf{M}^{*T} & \mathbf{T}^* \end{pmatrix}, \quad (10)$$

where we have $\mathbf{M}^* = \text{diag}[\mathbf{M}^{(1)}, ..., \mathbf{M}^{(m-1)}]^T$ and $\mathbf{T}^* = \text{diag}[\mathbf{T}^{(1)}, ..., \mathbf{T}^{(m-1)}]^T$. Finally, the N-R update consists of the following iterative formula:

$$\begin{pmatrix} \tilde{\mathbf{a}}_{(t+1)} \\ \tilde{\mathbf{\Psi}}_{(t+1)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{a}}_{(t)} \\ \tilde{\mathbf{\Psi}}_{(t)} \end{pmatrix} - \tilde{\mathbf{H}}^{-1}\tilde{\mathbf{g}}. \quad (11)$$

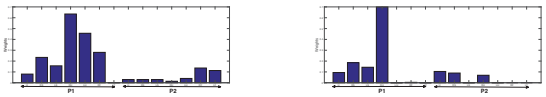
where $\tilde{\mathbf{a}} = [\mathbf{a}^{(1)^T}, \mathbf{a}^{(2)^T}, ..., \mathbf{a}^{(m-1)^T}]^T$ and $\tilde{\mathbf{\Psi}} = [\Psi^{(1)^T}, \Psi^{(2)^T}, ..., \Psi^{(m-1)^T}]^T$.

Experiments results

To evaluate the performance of our method, we conducted experiments on the UT-interaction dataset (Ryoo et al. 2009) which contains six classes of human-human interactions: *punch, kick, hand-shake, hug, points* and *push*. The dataset is divided into sets I end II, each one consisting of 10 videos for each activity. Set I has a static background and Set II is slightly more challenging with some camera motion. As porposed in (Ryoo et al. 2009), for two interacting persons we use for each activity the first four sequences from set I and the first three sequences from set II. This gives 24 and 18 instances from sets I and II, respectively, with an average number of 40 frames per video sequence. For each set, we randomly generated 5 groups for learning and 5 groups for testing and we average the obtained classification accuracy values.

Figure 3 shows weights obtained for *punch* and *point* interactions. From this figure we can note that our method has been able to attribute the highest weight for the trajectory which is discriminative for the interaction and lower weight values for trajectories having less discrimination.

Figure 4 shows the best results obtained using our method on the UT-dataset. We can note that generally our method outperforms the others for interaction classification. However, we can observe some classification errors between actions in each confusion matrix: between *hand shake* and *push* on Set1 confusion matrix, and confusion between *kick* and *push* on Set II confusion matrix. We compared our method to stat-of-art performance, and obtained classification accuracy are dressed in Table 1. (Liang et al. 2016) using spatio-temporal motion has achieved average accuracy of 84% and 92.3% with both spatio-temporal motion features and context features which does not reach the results that we have obtained 95.12%. These results demonstrate the performance of the proposed method even though we used a simple representation of the trajectories.



Punch *Point*

Figure 3: Weights obtained for punch and point interaction.

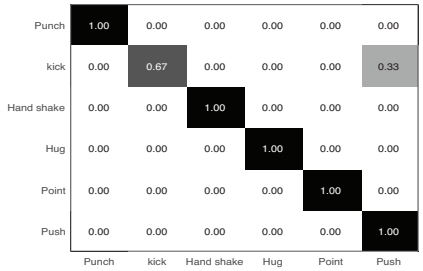
Conclusion

We proposed an approach for recognizing human interaction by introducing group of features weighting in multinomial kernel logistic regression. Our method assigns weights to joint trajectories depending on their contribution to discriminate between different interactions. Application of our method to the UT dataset has obtained results outperforming recent methods in terms of classification accuracy. In the future, we plan to extend our work with the use of complex representation such as deep learning networks.

Methods	Set I	Set II
Yuan et al. 2012	78.3	68.2
Yun et al. 2012	91.1	87.3
Meng et al. 2012	91.81	83.6
Sener et al. 2015	95	*91.67*
Motiian et al. 2017	*95.08*	89.39
Kernel logistic regression	87.5	83.33
Our method	**95.8**	**94.44**

Table 1: Average classification accuracy obtained by compared methods for set I and set II from UT-dataset.

Set I



Set II

Figure 4: Confusion matrix showing results in set I and set II from UT-interaction dataset.

Acknowledgment

The completion of this research has been made possible with the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Aggarwal, J. K., and Ryoo, M.S., Human Activity Analysis: A Review. *ACM Comp. Surveys*, 43(3), Article 16, 2011.

Datta, A., Shah M., da Vitoria Lobo, N., Person-on-Person Violence Detection in Video Data. *ICPR*, 433-438, 2002.

Hastie, T., Tibshirani, R., and Wainwright, M., *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

Liang, J., Xu, C., Feng, Z., and Ma, X., Affective Interaction Recognition Using Spatiotemporal Features and Context. *CVIU*, 144, 155-165, 2016.

C. Rao, A. Yilmaz, and M. Shah. View-Invariant Representation and Recognition of Actions. IJCV, 2002.

Meng, L., Qing, L., Yang,P. Miao, J., and Chen., X., Activity Recognition Based on Sementic Spatial Relation. *ICPR*, 609-612, 2012.

Motiian, S., Siyahjani, F., Almohsen, R., and Doretto, G., Online Human Interaction Detection and Recognition With Multiple Cameras. *IEEE Trans. on Circuits and Sys. for Video tech.*, 27(3):649-663, 2017.

Noceti, N., and Odone, F., Human in Groups: The Importance of Contextual Information for Understanding Collective Activities. *PR*, 47(11):3535-3551, 2014.

Ouyed, O., and Allili, M.S., Feature Weighting for Multinomial Kernel Logistic Regression and Application to Action Recognition. *ICPR*, 1325-1329, 2014.

Rao, C., Yilmaz, A., and Shah, M., View-Invariant Representation and Recognition of Actions. *IJCV*, 50(2), 203-226, 2002.

Park, S., and Aggarwal, J.K., Recognition of Two-person Interactions using a Hierarchical Bayesian Network. *ACM SIGMM Int'l Workshop on Video surveillance*, 65-76, 2003.

Do, T., and Pustejovsky, J., Learning event represrntation: As sparse as possible, but not sparser. arXiv preprint: arXiv:1710.00448v1, 2017.

Rao, N., Nowak, R., Cox, C., and Rogers, T., Classification With the Sparse Group Lasso. *IEEE Trans. on Signal Processing*, 64(2):448-463, 2016.

Ryoo, M.S., and Aggarwal J.K., Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities.*IEEE ICCV*, 1593-1600, 2009.

Schuldt, C., Laptev, I., and Caputo, B., Recognizing Human Actions: a Local SVM Approach. *ICPR*, 32-36, 2004.

Sener, F., and Cinbis, N.I., Two-Person Interaction Recognition via Spatial Multiple Instance Embedding. *J. of Visual Communication and Image Representation*, 32:63-73, 2015.

Slimani, K.N.E., Benezeth, Y., and Souami, F., Human Interaction Recognition Based on the Co-occurrence of Visual Words. *IEEE CVPR Workshops (CVPRW)*, 455-460, 2014

Tibshirani, R., Regression Srinkage and Selection via the Lasso. *J. of the Royal Statistical Society*, B, 58(1):267-288, 1996.

Vincent, M., and Hansen, N-R., Sparse Group Lasso and High Dimensional Multinomial Classification. *Computational Statistics & Data Analysis* 71:771-786, 2014.

Wang, J., Zhao, Z-Q.,Hu, X., Cheung, Y-M.Wang, M., and Wu, X., Online Group Feature Selection. *IJCAI*, 1757-1763, 2013.

Wang, H., and Leng, C., A Note on Adaptive Group Lasso. *Computational Statistics & Data Analysis*, 52(12):5277-5286, 2008.

Yuan, M., and Lin, Y., Model Selection and Estimation in Regression with Grouped variables. *J. of the Royal Statistical Society, Series B*, 68(1):49-67, 2006.

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., and Samaras, D., Two-person interaction detection using bodypose features and multiple instance learning. *IEEE CVPR Workshops*, 28-35, 2012.

Yuan, F., Xia, G-S., Sahbi, H., and Prinet, V., Mid-Level Features and Spatiotemporal Context for Activity Recognition. *Pattern Recognition*, 45(12):4182-4191, 2012.