Classifying Behavioral Attributes Using Conditional Random Fields

Michalis Vrigkas[1], Christophoros Nikou[1], and Ioannis A. Kakadiadis[2]

[1] Dept. of Computer Science & Engineering, University of Ioannina,
45110 Ioannina, Greece
[2] Computational Biomedicine Lab, Dept. of Computer Science,
University of Houston,
4800 Calhoun Rd, Houston, TX 77204, USA
{mvrigkas,cnikou}@cs.uoi.gr, ioannisk@uh.edu

Abstract. A human behavior recognition method with an application to political speech videos is presented. We focus on modeling the behavior of a subject with a conditional random field (CRF). The unary terms of the CRF employ spatiotemporal features (i.e., HOG3D, STIP and LBP). The pairwise terms are based on kinematic features such as the velocity and the acceleration of the subject. As an exact solution to the maximization of the posterior probability of the labels is generally intractable, loopy belief propagation was employed as an approximate inference method. To evaluate the performance of the model, we also introduce a novel behavior dataset, which includes low resolution video sequences depicting different people speaking in the Greek parliament. The subjects of the *Parliament* dataset are labeled as friendly, aggressive or neutral depending on the intensity of their political speech. The discrimination between friendly and aggressive labels is not straightforward in political speeches as the subjects perform similar movements in both cases. Experimental results show that the model can reach high accuracy in this relatively difficult dataset.

Keywords: Human Behavior Recognition, Conditional Random Field (CRF), Loopy Belief Propagation (LPB).

1 Introduction

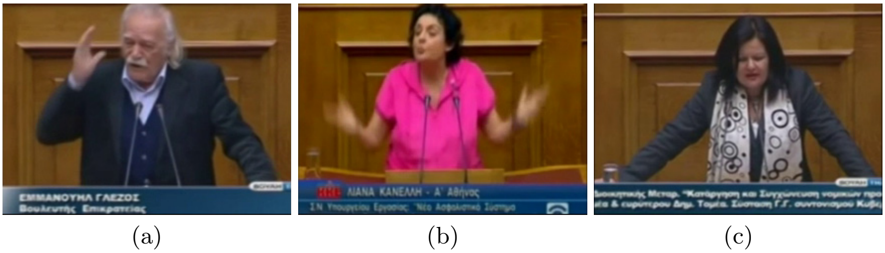
Recognizing human behaviors from video sequences is a challenging task for the computer vision community [1,10,13]. A behavior recognition system may provide information about the personality and psychological state of a person and its applications vary from video surveillance to human-computer interaction.

The problem of human behavior recognition is challenging for several reasons. First, constructing a visual model for learning and analyzing human movements is difficult. Second, the fine differences between similar classes and the short duration of human movements in time make the problem difficult to address. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. The variation of appearance, lighting conditions and

A. Likas, K. Blekas, and D. Kalles (Eds.): SETN 2014, LNAI 8445, pp. 95–104, 2014.

frame resolution makes the recognition problem amply challenging. Finally, the inadequate benchmark datasets are another obstacle that must be overcome.

In this paper, we are interested in characterizing human activities as behavioral roles in video sequences. The main contribution of this work is twofold. First, we introduce a method for recognizing behavioral roles (i.e., friendly, aggressive and neutral) (Figure 1). These behavioral classes are similar, as the involved people perform similar body movements. Our goal is to recognize these behavioral states by building a model, which allows us to discriminate and correctly classify human behaviors. To solve this problem, we propose an approach based on conditional random fields (CRF) [5]. Motivated by the work of Domke [2], which takes into account both model and inference approximation methods to fit the parameters for several imaging problems, we develop a structured model for representing scenes of human activity and utilize a marginalization fitting for parameter learning. Secondly, to evaluate the model performance, we introduce a novel behavior dataset, which we call the *Parliament* dataset [16], along with the ground truth behavioral labels for the individuals in the video sequences. More specifically, we have collected 228 low-resolution video sequences (320 × 240, 25fps), depicting 20 different individuals speaking in the Greek parliament. Each video sequence is associated with a behavioral label: friendly, aggressive and neutral, depending on the intensity of the political speech and the specific individual's movements.



(a) (b) (c)

Fig. 1. Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral.

The remainder of the paper is organized as follows: in Section 2, a brief review of the related work is presented. Section 3 presents the proposed approach including the model's specifications and the details of the method. In Section 4, the novel behavior recognition dataset is presented and experimental results are reported. Finally, conclusions are drawn in Section 5.

2 Related Work

The human activity categorization problem has remained a challenging task in computer vision for more than two decades. Many surveys [1,10] provide a good overview of human behavior recognition methods and analyze the properties of human behavior categorization. Previous work on characterizing human behavior has shown great potential in this area.

In this paper, the term “behavior” is used to describe both activities and events which are apparent in a video sequence. We categorize the human behavior recognition methods into two main categories: single- and multi-person interaction methods.

Single-Person Methods. Much research has focused on single person behavior recognition methods. A major family of methods relies on optical flow which has proven to be an important cue. Earlier approaches are based on describing behaviors by using dense trajectories. The work of Wang *et al.* [17] focused on tracking dense sample points from video sequences using optical flow. Yan and Luo [18] have also proposed an action descriptor based on spatio-temporal interest points (STIP) [7]. To avoid overfitting they have proposed a novel classification technique by combining the Adaboost and sparse representation algorithms. Our earlier work Vrigkas *et al.* [15] focused on recognizing single human behaviors by representing a human action with a set of clustered motion trajectories. A Gaussian mixture model was used to cluster the motion trajectories and the action labeling was performed by using a nearest neighbor classification scheme.

Multi-person Interaction Methods. Social interactions are an important part of human daily life. Fathi *et al.* [3] modeled social interactions by estimating the location and orientation of the faces of the persons taking part in a social event, computing a line of sight for each face. This information is used to infer the location an individual person attended. The type of interaction is recognized by assigning social roles to each person. The authors were able to recognize three types of social interactions: dialogue, discussion and monologue. Human behavior on sport datasets was introduced by Lan *et al.* [6]. The idea of social roles in conjunction with low-level actions and high-level events model the behavior of humans in a scene. The work of Ramanathan *et al.* [12] aimed at assigning social roles to people associated with an event. They formulated the problem by using a CRF model to describe the interactions between people. Tran *et al.* [14] presented a graph-based clustering algorithm to discover interactions between groups of people in a crowd scene. A bag-of-words approach was used to describe the group activity, while a SVM classifier was used to recognize the human activity.

3 Behavior Recognition Using CRF

In this paper, we present a supervised method for human behavior recognition. We assume that a set of training labels is provided and every video sequence is pre-processed to obtain a bounding box of the human in every frame and every person is associated with a behavioral label.

The model is general and can be applied to several behavior recognition datasets. Our method uses CRFs (Figure 2) as the probabilistic framework for modeling the behavior of a subject in a video. First, spatial local features are computed in every video frame capturing the roles associated with the bounding boxes. Then, a set of temporal context features are extracted capturing the relationship between the local features in time. Finally, the loopy belief propagation (LBP) [8] approximate method is applied to estimate the labels.

Let $r_j^t \in \mathcal{R}$ be the behavioral role label of the j^{th} person in a bounding box at frame t, where $\mathcal{R}$ is the set of possible behavioral role labels and $t \in [0, T]$ is the current frame. Let x_j^t represent the feature vector of the observed j^{th} person at frame t. Our goal is to assign each person a behavioral role by maximizing the posterior probability:

$$\mathbf{r} = \arg\max_{\mathbf{r}} p(\mathbf{r}|\mathbf{x}; \mathbf{w}). \tag{1}$$

It is useful to note that our CRF model is a member of the exponential family defined as:

$$p(\mathbf{r}|\mathbf{x}; \mathbf{w}) = \exp\left(E(\mathbf{r}|\mathbf{x}; \mathbf{w}) - A(\mathbf{w})\right), \tag{2}$$

where $\mathbf{w}$ is a vector of parameters, $E(\mathbf{r}|\mathbf{x})$ is a vector of sufficient statistics and $A(\mathbf{w})$ is the log-partition function ensuring normalization:

$$A(\mathbf{w}) = \log \sum_{\mathbf{r}} \exp\left(E(\mathbf{r}|\mathbf{x}; \mathbf{w})\right). \tag{3}$$

Different sufficient statistics $E(\mathbf{r}|\mathbf{x}; \mathbf{w})$ in (2) define different distributions. In the general case, sufficient statistics consist of indicator functions for each possible configuration of unary and pairwise terms:

$$E(\mathbf{r}|\mathbf{x}; \mathbf{w}) = \sum_j \Psi_u(r_j^t, x_j^t; w_1) + \sum_j \sum_{k \in \mathcal{N}_j} \Psi_p(r_j^t, r_k^{t+1}, x_j^t, x_k^{t+1}; w_2), \tag{4}$$

where $\mathcal{N}_j$ is the neighborhood system of the j^{th} person for every pixel in the bounding box. In our model temporal and spatial neighbors are considered. We use eight spatial and 18 temporal neighbors. The parameters w_1 and w_2 are the unary and the pairwise weights that need to be learned and $\Psi_u(r_j^t, x_j^t; w_1)$, $\Psi_p(r_j^t, r_k^{t+1}, x_j^t, x_k^{t+1}; w_2)$ are the unary and pairwise potentials, respectively.

Unary Potential: This potential predicts the behavior label r_j^t of the j^{th} person in frame t indicating the dependence of the specific label on the location of the person. It may be expressed by:

$$\Psi_u(r_j^t, x_j^t; w_1) = \sum_{\ell \in \mathcal{R}} \sum_j w_1 \mathbb{1}(r_j^t = \ell) \psi_u(x_j^t), \tag{5}$$

where $\psi_u(x_j^t)$ are the unary features and $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1, if the j^{th} person is associated with the ℓ^{th} label and 0 otherwise. The unary features are computed as a 36-dimensional vector of HoG3D values [4] for each bounding box. Then, a 64-dimensional spatio-temporal feature vector (STIP) [7] is computed, which captures the human motion between frames. The spatial relationship of each pixel in the bounding box and its 8×8 neighborhood is computed using a 16-dimensional Local Binary Pattern (LBP) feature vector [9]. The final unary features occur as a concatenation of the above features to a 116-dimensional vector.

Pairwise Potential: This potential represents the interaction of a pair of behavioral labels in consecutive frames. We define the following function as the pairwise potential:

$$\Psi_p(r_j^t, r_k^{t+1}, x_j^t, x_k^{t+1}; w_2) = \sum_{\substack{\ell \in \mathcal{R}, \\ m \in \mathcal{N}_\ell}} \sum_{\substack{j, \\ k \in \mathcal{N}_j}} w_2 \mathbb{1}(r_j^t = \ell) \mathbb{1}(r_k^{t+1} = m) \psi_p(x_j^t, x_k^{t+1}), \tag{6}$$

where $\psi_p(x_j^t, x_k^{t+1})$ are the pairwise features. We compute a 4-dimensional spatio-temporal feature vector, which is the concatenation of the 2D velocity and acceleration of the j^{th}person along time. The acceleration features play a crucial role in the distinction between the behavioral classes, as different persons in different behavioral classes perform similar movements. In addition, the L_2 norm of the difference of the RGB values at frames t and $t+1$ is computed. We use eight spatial and 18 temporal neighbors creating an 18-dimensional feature vector. The final pairwise features are computed as the concatenation of the above features to a 22-dimensional vector.

To learn the model weights $\mathbf{w} = \{w_1, w_2\}$, we employ a labeled training set and seek to minimize:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{\mathbf{r}} L(\mathbf{r}, \mathbf{x}; \mathbf{w}), \tag{7}$$

where $L(\cdot, \cdot)$ is a loss function, which quantifies how well the distribution (2) is defined by the parameter vector $\mathbf{w}$ matches the labels $\mathbf{r}$.

We select a clique loss function [2], which is defined as the log-likelihood of the posterior probability $p(\mathbf{r}|\mathbf{x}; \mathbf{w})$:

$$L(\mathbf{r}, \mathbf{x}; \mathbf{w}) = -\log p(\mathbf{r}|\mathbf{x}; \mathbf{w}). \tag{8}$$

The loss function is minimized using a gradient-descent optimization method. It can be seen as the empirical risk minimization of the Kullback-Leibler divergence between the true and predicted marginals.

Having set the parameters $\mathbf{w}$, an exact solution to Eq. (1) is generally intractable. For this reason, approximate inference is employed to solve this problem. In this paper, LBP [8] is used for computing the marginals using the full graphical model as depicted in Figure 2. For comparison purposes and for better insight of the proposed method, we have also tested a variant of the full graphical model by transforming it into a tree-like graph (Figure 3). This is accomplished by ignoring the spatial relationship between the observation nodes $\mathbf{x}$ and keeping only the temporal edges between the labels $\mathbf{r}$. In this case, tree-reweighted belief propagation [11] is considered for inference.

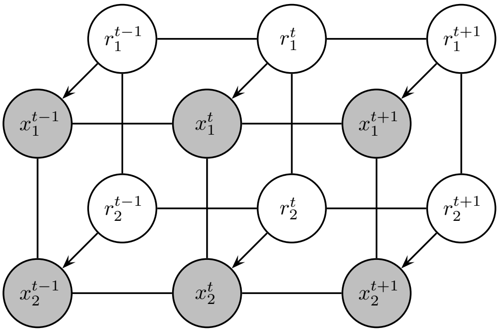


Fig. 2. Graphical representation of the model. The observed features are represented by $\mathbf{x}$ and the unknown labels are represented by $\mathbf{r}$. Temporal edges exist also between the labels and the observed features across frames.

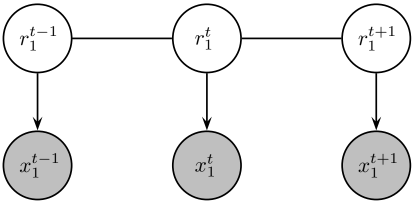


Fig. 3. Tree-like graphical representation of the model. The observed features are represented by $\mathbf{x}$ and the unknown labels are represented by $\mathbf{r}$.

4 Experiments

The experiments are applied to the novel *Parliament* dataset [16]. The number of features are kept relatively small in order not to increase the model's complexity. Additionally, to show that the proposed method can perform well, different model variants are compared.

To evaluate our method, we collected a set of 228 video sequences, depicting political speeches in the Greek parliament, at a resolution of 320×240 pixels at 25 fps. The video sequences were manually labeled with one of three behavioral roles: friendly, aggressive, or neutral, according to specific body movements. These behaviors were recorded for 20 different subjects. The videos were acquired with a static camera and contain uncluttered backgrounds. Figure 1 delineates some representative frames of the *Parliament* dataset.

We used 5-fold cross validation to split the dataset into training and test sets. Accordingly, the model was learned from 183 videos, while the algorithm was tested on the remaining five videos and the recognition results were averaged over all the examined configurations of training and test sets. Within each class, there is a variation in the performance of an action. Each individual exhibits the same behavior in a different manner by using different body movements. This is an interesting characteristic of the dataset which makes it quite challenging.

Table 1. Behavior classification accuracies (%) using the graphical model with only temporal edges (3) and the full graphical model (2) presented in Figure 2

Classification Accuracy(%)			
Method	**Friendly**	**Aggressive**	**Neutral**
Tree model (tree-reweighted BP)	100	49.23	84.48
Full model (loopy BP)	100	60.73	95.79

Table 2. Comparison between variants of the proposed method

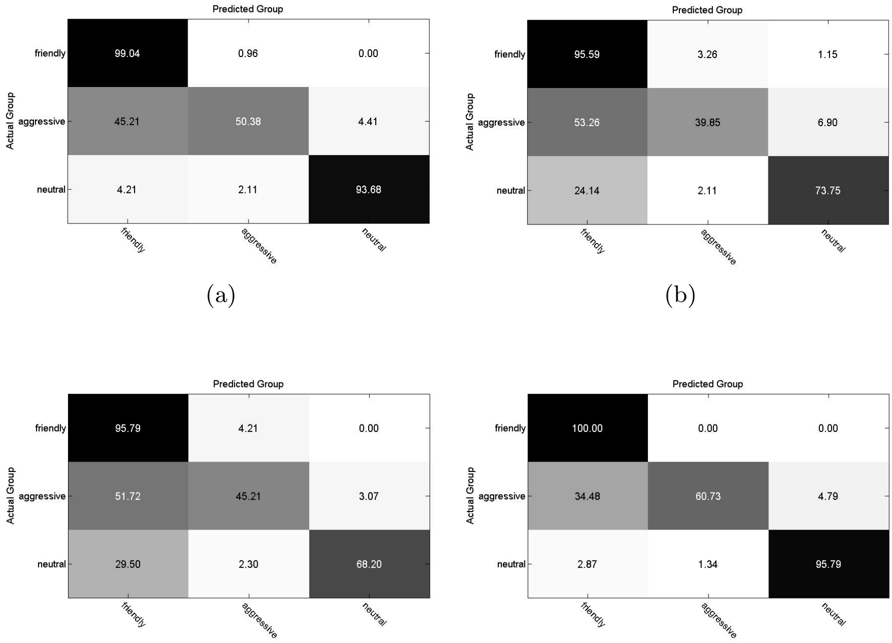
Method	**Accuracy(%)**
CRF (unary only)	81.0
CRF (unary no spatio-temporal)	69.7
CRF (pairwise no spatio-temporal)	69.7
Full CRF model	**85.5**

We evaluated the proposed model with different variants of the method. First, we compared the full graphical model (Figure 2) with a variant of the method, which considers the graphical model as a tree-like graph (Figure 3). As it can be observed in Table 1, the full graphical model performs better than the tree-like graph, which uses only temporal edges between the labels. The second model ignores the spatial relationship between the features and the classification error is increased. Generally, the full graphical model provides strong improvement of more than 8% with respect to the tree model.

In the second set of experiments, we evaluated three variants of the proposed CRF model. First, we used the CRF model with only the unary potentials ignoring the pairwise potentials. The second variant uses only unary potential without the spatio-temporal features. Finally, the third configuration uses the full model without the spatio-temporal pairwise features. The classification results comparing the different models are shown in Table 2.

We may observe that the CRF model, which does not use spatio-temporal feature in either the unary potentials or the pairwise potentials, attains the worst performance between the different variants. It is worth mentioning that the first variant, which uses only unary features, performs better than the other two variants, which do not use spatio-temporal features. However, this is not a surprising fact, as in the case of the no spatio-temporal variants the classification is performed for each frame individually ignoring the temporal relationship between consecutive frames. The use of spatio-temporal features appears to lead to better performance than all the other approaches. We also observe that the full CRF model shows significant improvement over all of its variants. The full CRF model leads also to a significant increase in performance of 85.5%, with respect to the model with no spatio-temporal features. This confirms that temporal and spatial information combined together constitute an important cue for action recognition.

Figure 4 illustrates the overall behavior recognition accuracy, where the full CRF model exhibits the best performance in recognizing each of the three behaviors. The main conclusion we can draw from the confusion matrices is that adding temporal edges to the graphical model helps reduce the classification error between the different behavioral states. It is also worth noting that, due to missed and relatively close features in consecutive frames, the classes "friendly" and "aggressive" are often confused as the subject performs similar body movements. Feature selection may be employed to solve this problem.



(a) (b)

(c) (d)

Fig. 4. Confusion matrices of the classification results for the CRF model employing (a) only unary potentials, (b) only unary potentials without spatio-temporal features, (c) the full model without spatio-temporal pairwise features, and (d) the full model

5 Conclusion

In this paper, we have presented a method for recognizing human behaviors in a supervised framework using a CRF model. We have also introduced a new challenging dataset (*Parliament*), which captures the behaviors of some politicians in the Greek parliament during their speeches. Several variants of the method were examined reaching an accuracy of 85.5%.

A direction of future research would be to study how the use of voice features and pose can help improve recognition accuracy. We are also interested in studying feature selection techniques to better separate the classes “friendly” and “aggressive”. With these improvements, we plan to apply this method to several other datasets.

Acknowledgments. This work was supported in part by the University of Houston Hugh Roy and Lillie Cranz Cullen Endowment Fund.

References

1. Candamo, J., Shreve, M., Goldgof, D.B., Sapper, D.B., Kasturi, R.: Understanding transit scenes: A survey on human behavior-recognition algorithms. IEEE Transactions on Intelligent Transportation Sysstems 11(1), 206–224 (2010)
2. Domke, J.: Learning graphical model parameters with approximate marginal inference. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(10), 2454–2467 (2013)
3. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, USA, pp. 1226–1233 (2012)
4. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: Proc. British Machine Vision Conference, University of Leeds, Leeds, UK, pp. 995–1004 (September 2008)
5. Lafferty, J.D., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conference on Machine Learning, Williams College, Williamstown, MA, USA, pp. 282–289 (2001)
6. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, USA, pp. 1354–1361 (2012)
7. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2-3), 107–123 (2005)
8. Murphy, K.P., Weiss, Y., Jordan, M.I.: Loopy belief propagation for approximate inference: An empirical study. In: Proc. Uncertainty in Artificial Intelligence, Stockholm, Sweden, pp. 467–475 (1999)
9. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
10. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
11. Prince, S.J.D.: Computer Vision: Models Learning and Inference. Cambridge University Press (2012)
12. Ramanathan, V., Yao, B., Fei-Fei, L.: Social role discovery in human events. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, OR, USA (June 2013)
13. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 99(PrePrints), 1 (2013)

14. Tran, K.N., Bedagkar-Gala, A., Kakadiaris, I.A., Shah, S.K.: Social cues in group formation and local interactions for collective activity analysis. In: Proc. 8th International Conference on Computer Vision Theory and Applications, Barcelona, Spain, pp. 539–548 (February 2013)
15. Vrigkas, M., Karavasilis, V., Nikou, C., Kakadiaris, I.A.: Action recognition by matching clustered trajectories of motion vectors. In: Proc. 8th International Conference on Computer Vision Theory and Applications, Barcelona, Spain, pp. 112–117 (February 2013)
16. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: The Parliament database (2014), `http://www.cs.uoi.gr/~mvrigkas/Parliament.html`
17. Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Action recognition by dense trajectories. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, United States, pp. 3169–3176 (2011)
18. Yan, X., Luo, Y.: Recognizing human actions using a new descriptor based on spatial-temporal interest points and weighted-output classifier. Neurocomputing 87, 51–61 (2012)