REAL-LIFE VIOLENT SOCIAL INTERACTION DETECTION

Paolo Rota[1,2], *Nicola Conci*[1], *Nicu Sebe*[1], *James M. Rehg*[3]

University of Trento[1], TU Wien[2], Georgia Institute of Technology[3]

ABSTRACT

This paper proposes a method to detect and localize dyadic human interactions in real videos. The idea stems from the significant difference between an action performed by a single subject and an interaction between two persons. In the first case all the visual information is concentrated on the subject, while in the latter case the action of a person is related to the interacting person's attitude, following an action/reaction principle. This kind of behavior is significant especially in natural and real scenarios, in which people are moving freely without the awareness of being recorded. To highlight these features and provide researchers with a common ground for comparisons, we have collected and annotated a new dataset, retrieving from YouTube 30 different videos of a specific type of interaction, namely urban fight situations. The proposed dataset is one of the most challenging annotated video collection concerning dyadic interactions, due to the intrinsic intra-class variability characterizing real fights. In addition, we provide an extensive experimental analysis on this dataset and we demonstrate that the visual information extracted in the area associated to the interpersonal space plays a fundamental role in detecting fights.

Index Terms— Fight Detection, Real-Life Scenario, Dataset, Scene Analysis.

1. INTRODUCTION

In today's digital age, the enhancement of the hardware technology has set new horizons on the computer vision universe, fostering researchers to ask new questions, tackling problems and finding new solutions. The research in video analysis, in the last years, has proposed significant improvements in action/activity recognition [1, 2, 3, 4]. The huge interest in this field as well as the performance improvement in detecting and analyzing visual features extracted from videos, have pushed researchers to consider also situations involving more than one person [5, 6, 7, 8].

Social interactions have been investigated mostly in controlled scenarios [9, 4, 10] ignoring the fact that in real-life persons may be hardly detectable, illumination may vary rapidly altering the camera perception, people may perform the same action in many different ways, etc. These reasons led us to consider that the real life scenario is a further and inevitable step to be done in order to bring social interaction analysis a bit closer to an applicative level.

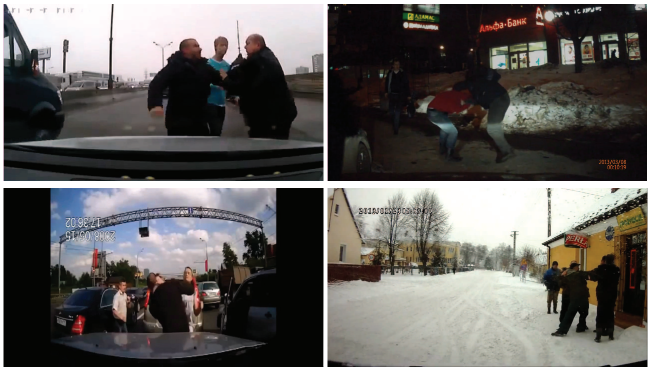


Fig. 1. Snapshots taken from the proposed dataset. We intend to emphasize the intra-class difference in fight interactions even if all samples are captured in the similar urban scenarios.

Another important element proposed in this work is the concept of unstructured interaction. In the majority of the datasets proposed so far [9, 11, 3], social physical interactions have been considered short and well defined. Let us take hand-shaking as an example: two people approach each other, move their arms forward, shake their hands, and move back to the original position. These sub-activities have unique meaning, no matter what the situation is or who the subjects involved are. The same could be for hugging, kissing, or other similar interactions. In this work we are interested in such kind of mutual actions that do not exhibit a predefined structure.

In order to emphasize this aspect we propose a new dataset of videos retrieved from YouTube (some snapshots are depicted in Fig. 1) with different kinds of physical unstructured social interactions. A particular type of interaction that matches the proposed characteristics is *urban fighting*. In fact in some public datasets [12, 9, 17, 13] there are several instances of fighting but they are usually staged and it is always clear that occasional actors are not performing naturally. Another aspect is the length of the events in the existing datasets. Since their purpose mainly regards events classification, the length of every execution is below 5 seconds. Our dataset is composed by *urban fights* in which the interaction lasts for more than 10 seconds (in some cases even minutes) and they often include more fighting instances in a single video.

In terms of detection of violent situations in videos, an

Table 1. A comparison among most popular datasets for social interaction analisys

Dataset	Number of sequences	Resolution	Scenario
UT-Interaction Dataset [9]	20	720x480	Staged
Caviar Test Case Scenario [12]	28	384x288	Staged
BEHAVE Interactions Test Case Scenario [13]	8	640x480	Staged
Collective Activity Dataset [14]	74	Variable	Natural/staged
UCLA Courtyard Dataset [15]	6	2560X1920	Natural
Hockey Fights Dataset [16]	1000	360x288	Single Scenario
TVHI dataset [11]	300	Variable	Movies
Hollywood2 [3]	2517 (771 are interactions)	Variable	Movies
RE-DID (Our Dataset)	**30**	**1280X720**	**Natural**

early work [18] proposes the analysis combining different visual (blood, flames, etc.) and audio features (explosions, screams, etc.). Chang et al. [19] propose a multi-camera framework to detect and predict aggressive behaviors between groups of individuals such as gangs in prison yards; in their model each individual is tracked along the monitored area. They propose a hierarchical clustering to define groups of individuals in order to detect predefined behaviors such as loitering, flanking, and aggressive group behaviors. Nievas et al. [16] propose a classification problem to detect hockey fights in short video clips collecting visual features over the whole frame. Hassner et al. [20] propose a real-time detection model of violent crowd behaviors using flow information.

In our approach we focus our attention on dyadic aggressive interactions, so we are not evaluating group aggressions as in [19]; on the other hand, unlike [16], we focus our attention on a restricted interpersonal space collocated between the two opponents in order to prune the visual features not related to the ongoing event. Our final goal is also to detect the fight situations on-line as soon as they happen in the video.

As a further contribution, a novel method to detect and localize pairwise unstructured physical social interaction is proposed. The method relies on the definition of an interpersonal area between the interacting subjects, in which the motion cues are intense and therefore are more discriminative. As a further motivation, the interpersonal space includes inherently the proxemic information among the interacting subjects, providing an important contribution that can not be provided by the visual features alone.

The paper is organized as follows. In Section 2 we present in detail our new dataset, while in Section 3 we propose our evaluation methodology. In Section 4 we present the detection results, while conclusions are drawn in Section 5.

2. REAL-LIFE EVENTS - DYADIC INTERACTIONS DATASET (RE-DID)

A crucial contribution given by this paper consists on the collection of a new dataset for dyadic interactions called Real-life Events Dyadic Interaction Dataset (ReDID). The main motivation that pushed us to propose a new dataset is given by the lack of annotated videos recorded in real-life scenarios, picturing unstructured challenging interactions performed by a pair of subjects.

To support our claim, Tab. 1 shows some characteristics of the most popular datasets published so far. To be more specific, the data proposed by [9, 12, 13] are more focused on dyadic/small group interactions but the number of situations is smaller than ours and, more important, videos are staged so there is no spontaneity in performing actions. Datasets [14, 15] are more dedicated to group interaction and single action recognition; they also lack unstructured interactions that are substantial in our work. The dataset proposed by Nieves et al. [16] is composed by very short clips (50 frames each) taken from ice hockey games including fight scenes and normal game situations. The dataset has a retrieval/classification purpose, the fights are taken from the same scenario, and there is no huge variability in the way fights take place. Additionally, these fights are far from the violent situations occurring in a surveillance context, which instead are addressed in our dataset. Other datasets [11, 3] are more dedicated to classification/retrieval, since they are composed of a higher number of videos. Moreover, videos in these datasets are short and often taken from movies. In [13, 12, 9] fighting situations are present but their dynamics is often inconsistent compared to what happens in real situations, and the number of examples is not large enough to obtain reliable statistics.

All the videos in the dataset are retrieved from YouTube; 25 of them are recorded using car mounted Dash-Cams, the remaining ones have been taken by other devices such as mobile phones. The length of the videos varies from 0:20 to 4:02 (mm:ss) and the resolution has been normalized to 1280x720 for the sake of homogeneity. The dataset includes 73 different fight instances under different lighting (day, night) and weather conditions (sunny, rainy), different original video resolution (native 1280x720, upsampled videos), different camera views (wide angle, fish-eye, zoomed view), moving and static scenes.

The dataset has annotations of the position of the subjects' bounding boxes for each frame and relative ID, the temporal window where the interaction occurs, and the position of the interpersonal spaces (see the definition in section 3) precomputed for the ground truth. For what concerns the interaction triggering and ending, we have considered a general rule

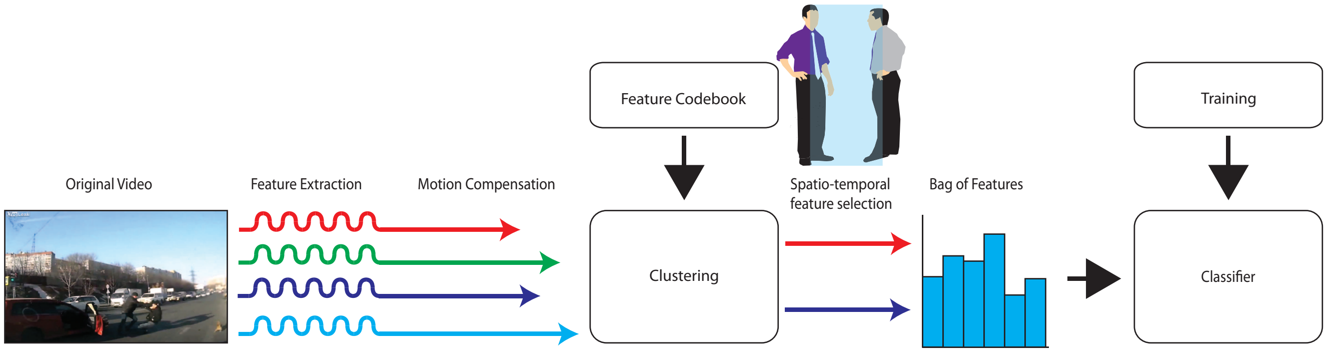


Fig. 2. The framework schema for the fight detection evaluation using the interpersonal space.

for the annotation process, starting with the first contact between the involved subjects until a relevant distancing is takes place.[1]

3. EVALUATION FRAMEWORK

In this section we describe the method used to evaluate our framework for social interactions analysis. Fig. 2 shows a global overview of the approach we propose.

In order to capture both shape and motion features, dense trajectories have been adopted. The extraction of the dense trajectories is affine with the proposal by Wang et al. [21]. The interesting points are sampled using Shi and Tomasi algorithm [22] and then tracked using the Farnebäck's implementation [23] of the optical flow. Most of the videos in Re-DID are affected by high camera motion, that it is fairly fluid in the case of Dash-Cam videos, but that often degrades in mobile recorded clips into rapid shaking. To improve the cleanness of the trajectories we apply a homographic correction according to [24]. The mean coordinate of the points in the trajectory has been considered as the location point for the feature extracted. The descriptors used in this work are essentially shape features i.e. HOG (Histogram Of Gradients), zero-order and first-order motion features as HOF (Histogram of Optical Flow), and MBH (Motion Boundary Histograms) [25, 26].

We evaluate social interactions according to two different criteria. The first is the *high-level* approach, in which trajectories of subjects are the most informative source. The second criterion is similar to the one used to classify atomic actions at the pixel-level information, thus analyzing color, gradient, and optical flow information. In fact, the proximity information between subjects and the analysis of the visual features created by the movement of the individuals, can equally contribute to provide an accurate description of the ongoing interaction. To support this claim it is enough to think about the situation when two people are interacting within a close-range distance, in which most of the salient motion occurs in the space located between them.

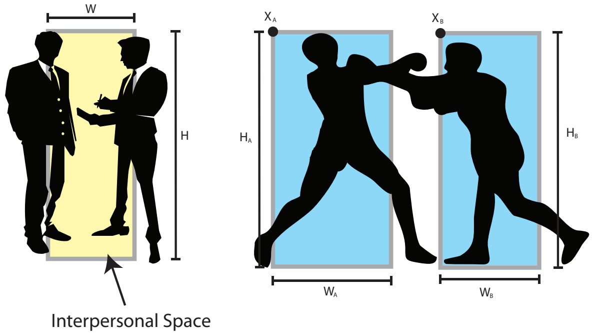


Fig. 3. The sketch shows two situations of non-fight and fight, respectively, with the notations used in Section 3 to locate the *interpersonal space* position.

Let U_a and U_b be the collection of shape and position parameters referred to subject a and b respectively as depicted in Fig. 3. Let's define what we call *interpersonal space* through the following equations:

$$H = \max\left(H_a, H_b\right) \tag{1}$$

$$W = |X_a - X_b| \tag{2}$$

$$K_w \cdot \min(W_a, W_b) \leq W \leq W_a + W_b \tag{3}$$

$$\left|\frac{H_a}{H_b} - 1\right| \leq K_h \tag{4}$$

In Eq. (1) and Eq. (2) we define the height and the width of the interpersonal bounding box, respectively. In Eq.(3) we define the dimension constraints (K_w and K_h are two constant values that depends on interacting subject's aspect rateo). The first part of the equation prevents the minimum area to be smaller than the narrowest person's bounding box in order to manage the situation in which the involved subjects are occluding each other. The second part is a distance constraint; whenever it is not satisfied we assume that every close-range interaction is not possible. Eq.(4) is a perspective constraint that avoids considering as interacting two subjects that are too far from each other. On top of these considerations the interpersonal space is always centered onto the center of the conjunction line between the two subjects' bounding boxes.

Feature extraction has been pursued according with Eq. (5) where $\phi_{ab}(t)$ are the extracted features related to

[1] The dataset is available at `http://mmlab.science.unitn.it/ReDID/`

Table 2. Results for Fight localization on UT Dataset and ReDID.

	AUC [%]				
	UT-Dataset		**Re-DID**		
	Spatial no tracker	**Cuboid no tracker**	**Spatial no tracker**	**Cuboid no tracker**	**Spatial tracker**
HOG+HOF					
P	75.63	82.36	63.52	67.71	63.55
I	82.31	87.01	71.21	73.26	75.87
P+I	81.77	84.51	71.68	71.22	73.59
MBH					
P	76.43	83.77	65.45	64.12	67.43
I	82.82	88.84	72.14	72.46	76.18
P+I	82.23	88.39	74.10	70.31	74.73
HOG+HOF+MBH					
P	78.93	84.38	64.94	68.72	65.74
I	83.41	92.25	72.70	71.33	71.50
P+I	83.54	88.51	72.09	72.33	73.96

subject a and b at time t. $\boldsymbol{x}(t)$ is the whole set of N valid trajectories present at time t.

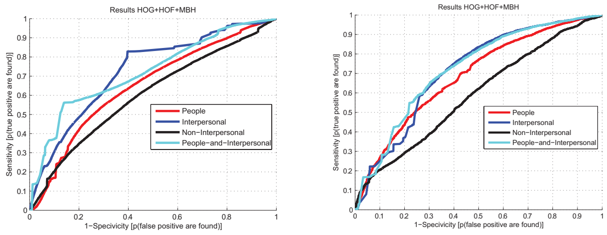
$$\phi_{ab}(t) = \{\boldsymbol{x}_i(t)\}_{i=1}^{N} : \boldsymbol{x}_i \in I_{ab}(t) \tag{5}$$

We indicate with $I_{ab}(t)$ the area of the interpersonal space at time t generated by U_a and U_b.

Only trajectories that lie in the interpersonal space are considered. The experiments we propose have two goals: the first is in the spatial domain, the second consists of collecting the data from a temporal cuboid similar to [27], but unlike them we did not collect data from subvolumes in a uniform grid. In this work the cuboid is defined by the temporal envelope of the interpersonal space bounding box along a predefined Δt. This modification has been necessary since we intend also to localize the interaction. The features are gathered inside the spatio-temporal envelope and processed using a bag of features approach. All the combinations among the subjects are considered but only where the interpersonal space is defined. If the conditions in Eq. (3)-(4) are satisfied we perform the actual classification based on a linear Support Vector Machine.

4. RESULTS

The results are presented in Tab. 2. The first part refers to the detection accuracy on the UT-Dataset [9]. In order to fit the videos to our task we put together some routine social interactions as hugging, pointing and handshaking, considering them as normal types of behaviors. On the other hand we gather pushing, kicking and punching as violent interactions. The second part of Tab. 2 refers to Re-DID, with an additional experiment using a well known tracker [28] to track automatically pedestrians in the videos. As an important remark, given the low detection rate of the tracker (less than 20%) due to the complexity of the scenario, we have reported the results of the binary classification only on the detected people and not on the overall cases.



(a) (b)

Fig. 4. Examples of ROC curves for frame-by-frame fight detection with (b) and without (a) temporal model on Re-DID.

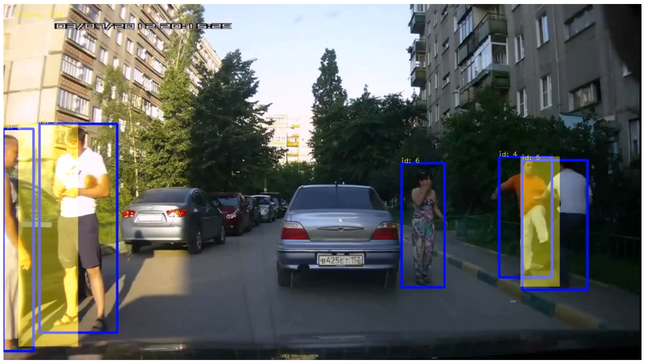


Fig. 5. An example of the discriminative capability of the interpersonal space model. In the figure it is highlighted how the two interacting subjects on the left are separated from the subjects on the right, moreover on the far right the two subjects are fighting while the one slightly on their left keeps distance and indeed is not considered as interacting with the previous two.

The benefit given by the introduction of the interpersonal space is evident in Tab. 2 and then confirmed by the ROC curves in Fig. 4. The experiments refers to the feature extraction on the subject's bounding box P, on the interpersonal space I and on both $P + I$. In the ROC curves we have also reported an extra experiment to prove our concept in which we collect $S : A - (P + I)$, where A is the whole set of features present in the frame. An example of fight detection in urban scenario using the proposed strategy is depicted in Fig. 5.

5. CONCLUSION

Addressing dyadic violent interactions in a real life scenario turns out to be a harder undertaking comparing to the traditional action/interaction analysis that is mostly performed on staged videos. The new dataset proposed in this paper aims to address this shortcoming of the currently available literature. According to the results presented in Sec. 4 we can confirm that proxemic information introduced by the interpersonal space is beneficial in terms of detection of the interaction itself and moreover in the discrimination among different types of behaviors.

6. REFERENCES

[1] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *ICCV*, 2005.

[2] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *ICPR*, 2004.

[3] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid, "Actions in context," in *CVPR*, 2009.

[4] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori, "Discriminative latent models for recognizing contextual group activities," *PAMI*, pp. 34.8: 1549–1562, 2012.

[5] Nuria M Oliver, Barbara Rosario, and Alex P Pentland, "A bayesian computer vision system for modeling human interactions," *PAMI*, pp. 22.8: 831–843, 2000.

[6] Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid, "High five: Recognising human interactions in tv shows," *BMVC*, 2010.

[7] Minh Hoai and Andrew Zisserman, "Talking heads: Detecting humans and recognizing their interactions," *CVPR*, 2014.

[8] De-An Huang and Kris M Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *ECCV*. 2014.

[9] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," `http://cvrc.ece.utexas.edu/SDHA2010/Human\_Interaction.html`, 2010.

[10] Wongun Choi, Khuram Shahid, and Silvio Savarese, "Learning context for collective activity recognition," in *CVPR*, 2011.

[11] Alonso Patron-Perez, Marcin Marszalek, Ian Reid, and Andrew Zisserman, "Structured learning of human interactions in tv shows," *PAMI*, pp. 34.12: 2441–2453, 2012.

[12] CAVIAR, "Test case scenario," `http://groups.inf.ed.ac.uk/vision/CAVIAR/`, 2004.

[13] BEHAVE, "Interactions test case scenarios," `http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/index.html`, 2007.

[14] Wongun Choi, Khuram Shahid, and Silvio Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *ICCV Workshops*, 2009.

[15] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *ECCV*, 2012.

[16] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*, 2011.

[17] Paolo Rota, Nicola Conci, and Nicu Sebe, "Real time detection of social interactions in surveillance video," in *ECCV Workshops and Demonstrations*, 2012.

[18] Jeho Nam, Masoud Alghoniemy, and Ahmed H Tewfik, "Audio-visual content-based violent scene characterization," in *ICIP*, 1998.

[19] Ming-Ching Chang, Nils Krahnstoever, Sernam Lim, and Ting Yu, "Group level activity recognition in crowded environments across multiple cameras," in *AVSS*, 2010.

[20] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *CVPR Workshop*, 2012.

[21] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.

[22] Jianbo Shi and Carlo Tomasi, "Good features to track," in *CVPR*, 1994.

[23] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*. Springer, 2003.

[24] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[25] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[26] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, 2006.

[27] Michael Sapienza, Fabio Cuzzolin, and Philip Torr, "Learning discriminative space-time actions from weakly labelled videos," in *BMVC*, 2012.

[28] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR*, 2011.