SCISPACE
formerly Typeset

Open access • Proceedings Article • DOI:10.1109/SMC.2014.6974283

Context-based Trajectory Descriptor for Human Activity Profiling — Source link

Eduardo M. Pereira, Lucian Ciobanu, Jaime S. Cardoso

Published on: 01 Oct 2014 - Systems, Man and Cybernetics

Topics: Context (language use)

Related papers:

- A new feature descriptor for 3D human action recognition
- Wavelet-based holistic sequence descriptor for generating video summaries
- Human action recognition using time-invariant key-trajectories describing spatio-temporal salient motion
- Tri-level Unified Framework for Human Gait Analysis
- Action Recognition with Temporal Relationships

Share this paper:

View more about this paper here: https://typeset.io/papers/context-based-trajectory-descriptor-for-human-activity-whsp4cesei

Context-based Trajectory Descriptor for Human Activity Profiling

Eduardo M. Pereira
INESC TEC and
Faculty of Engineering
of the University of Porto
Rua Dr. Roberto Frias, 378
Porto, Portugal 4200 - 465
Email: ejmp@inescporto.pt

Lucian Ciobanu
INESC TEC
Rua Dr. Roberto Frias, 378
Porto, Portugal 4200 - 465
Email: lucian.ciobanu@inescporto.pt

Jaime S. Cardoso
INESC TEC and
Faculty of Engineering
of the University of Porto
Rua Dr. Roberto Frias, 378
Porto, Portugal 4200 - 465
Email: jaime.cardoso@inescporto.pt

***Abstract*—The increasing demand for human activity analysis on surveillance scenarios has been provoking the emerging of new features and concepts that could help to identify the activities of interest. In this paper, we present a context-based descriptor to identify individual profiles. It accounts with a multi-scale histogram representation of position-based and attention-based features that follow a key-point trajectory sampling. The notion of profile is expressed by a new semantic concept introduced as an adjective for action recognition. We also identify a very rich dataset, in terms of intensity and variability of human activity, and extended it by manual annotation to validate the introduced concept of profile and test the descriptor's discriminative power. High rates of recognition were achieved.**

I. Introduction

Automatic behavior understanding from video is a very complicated problem. It comprises several hierarchical layers of processing, from bottom low-level features to top high-level semantics interpretation. The reduction of this gap is still a challenge in some type of applications, specially due to temporal domain. Mid-level descriptors intend to robustly represent spatiotemporal relationships between features and objects to discriminatively detect actions, events, and form atomic elements of a complex activity. This work addresses these problems by proposing a descriptor that considers relational information between an individual and the scene, namely objects of interest. Such contextual features are sample over a key-point trajectory scheme for multiple scales.

Spatiotemporal trajectory representations are gaining increasing attention on surveillance scenarios to analyze human activity and detect abnormal events. Indeed, the research community has been focusing on solving technical problems associated to multi-tracking techniques and on encoding trajectory-based features to detect individual atomic actions. Some approaches combine scene and object features with trajectory-based descriptors to detect event primitives [1], while others aggregate interactions measures and cues to analyze small groups of pedestrian [2]. However, none of them explore the integration of scene objects with individual related features to form action profiles. We formulate a relational descriptor that accounts with a trajectory-based motion component, position-based and attention-based features to build individual profiles, which can further describe semantic action phrases.

Our descriptor is build upon a bag-of-features methodology. Such descriptor considers a concatenation of histograms of relational and single spatiotemporal features taken from each trajectory's keypoint. This approach was tested on real scenarios with a proper annotation information, which gives a clear idea of its performance and potentiality. In fact, since we use annotated trajectories, we could evaluate its discriminative power and infer robust individual feature importance. In addition, we extend an existing dataset with low-level detection and tracking information, and with new high-level semantic concepts that encompass higher abstraction terms for action context. We also show that those concepts could be adapted to semantic annotations that had been presented in the literature.

II. Related Work

Trajectory dynamics provide intrinsic features that can be used to build useful representations to analyze several application-driven interests such as scene topology, event detection, social interpretation, and activity classification.

A common practice is to use trajectory information to model a scene by a topographical map composed of nodes, which are the areas of interest, and edges, which represent the connectivity between those areas and encode the activity of a human. The work of [3] classify the areas of interest as entry/exit zones, junctions, intersections, and stop areas, which are defined by trajectories characteristics. In [1] trajectory slow points define individual topologies which are combined to form the general topology. After that, they segment trajectories by topology affinity building a descriptor composed by primitive events. In general, the statistical and geometric structures inferred from the scene model could be used in a feedback loop, for instance they could filter out false detections or could enrich tracking approaches that incorporate scene context.

Trajectory information could help to detect typical and unusual events. In [4] the authors applied self-organising feature map neural networks, with trajectories encoded as point-based flow vectors, to learn normal trajectories and detect new trajectories event-related. However, it can not distinguish between new normal paths and abnormal behaviors. In [5] they solved this problem and stated an accuracy improvement for trajectory classification and event detection when Fourier coefficient space is used instead of trajectory space. However, tests were just carried on synthetic and manual annotated data, and as complex trajectories are not suitable to a global Fourier approximation, probably this approach would not perform as good on real scenarios.

The sequence of trajectories' characteristics are normally used to extract motion patterns that segment the scene into semantic regions. In [6] was introduced a clustering algorithm that accounts with similarity measures and comparison confidence between trajectories to obtain clusters of different activities. This type of approaches largely depends on the distance measures, which also vary depending on the activities being detected. To overcome those errors, some approaches formulate the problem in a probabilistic way. [7] proposed a nonparametric bayesian framework. The number of clusters for both the observations of an object on a trajectory and the trajectories are simultaneously learned from a dual hierarchical Dirichlet process (HDP). It also permits to reduce the space complexity. However, since trajectories are modeled as words to be quantized into a codebook, such representation lacks of temporal information.

An extension of activity analysis embeds notions of social psychology, normally applied to discover and characterize groups of people. Relational connections among people, focus of attention of each person, geometric scene constraints, and proxemics-based distances are ones of the cues used on this type of analysis. Normally, such information is used by microscopic approaches, whose can be divided into social force model-based [8], virtual agents [9], and cellular automata [10]. In particular, the work presented in [11] adopted a probabilistic grouping strategy which accounts with a pairwise spatiotemporal measure between people. A connectivity graph is built for further segmentation of groups and derivation of individual probabilistic models. Each model considers motion type, related to atomic action, direction distribution and distance change, related to interactions. However, no object-scene relation is considered, and they did not intend to use relational context to describe individual profiles. In fact, common microscopic approaches try to simulate pedestrian physical behavior and infer characteristics about group formation, dispersion, and evolution, but they do not capture individual semantics.

This paper shows three clear contributions: i) a new definition of semantic concepts, whose abstraction meaning intends to encompass existing semantic labels, normally used for atomic actions and event primitives, and form more complex visual phrases; ii) annotation of low-level and high-level information on a very rich dataset for human activity and a baseline to detect and classify individual profiles, both will be released to be used by the research community; iii) a relational context descriptor that accounts with spatiotemporal information from trajectory-based features and attention-based scene cues to robustly discriminate among several profiles.

III. Semantic Concepts and Annotation

Annotation of human nonverbal behavior should reveal meaningful representations that could be semantically associated with ontological concepts for human activity. The diversity of theories that intent to explain the linkage between the psychophysiological states and the human behavior has been triggering different representation approaches on the literature that account with temporal process of actions, spatiotemporal relationships between entities, poses, gestures, among others.

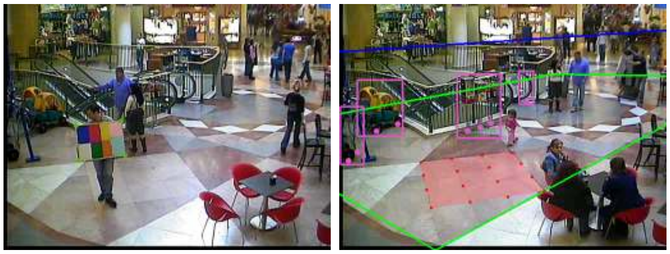
A general view about human activity analysis was presented in [12]. They defined a hierarchical approach where semantic levels are related to an increasing complexity of human activity categorization: i) gestures, elementary movements of body parts such as *raising an arm*; ii) actions, atomic activity composed by temporal sequence of gestures such as *jumping*; iii) interactions, a sequence of single activities between persons such as *a person hugging another*; iv) group activities, single or complex activities performed by a conceptual group such as *a group having a dinner*. Such levels follow an analogy to grammar-based semantics that can be used to map annotation labels to relational inferring models, for instance an action is associated to a verb, a gesture to a phrase where the entity is the body part, etc. Some works had already defined semantic concepts: division between the part of actions as objects and the poselets closely related to those actions [13], Allen's temporal predicates applied to features and entities to model activities with complex structure [14], definition of topological and directional relations between persons to build context-free grammars [15] and collective context descriptors [16].

Our aim is to add semantics to profile individual behaviors, a topic not so well explored on the literature. Following the grammar-based analogy, we present a deeper abstraction layer that can be associated to adjectives, since we are qualifying person's behavior characteristics. We look for a real-life surveillance dataset with large duration, intense activity, and high diversity of semantics in terms of individual and collective activity, in order to extend it for new human activity analysis challenges. We found the IIT (Israel Institute of Technology) dataset and grant permissions from the authors [17].

The dataset is composed by several urban scenarios such as shopping, subway, and street. We chose the shopping scenario since the social context provides more well-defined profiles. This scenario comprises three videos (resolution 512×384 @25 fps) with duration: 83155 frames (55‘26“), 59969 frames (39‘58“), and 90525 frames (60‘21“). Until this moment, due to hard manual labor, we only use the first one. We were advised by the lab of social-psychology of the University of Porto[1] during the annotation process. In specific, they help us to analyze and identify individual and collective profiles. Such identification was done with the knowledge of social influence, perception and interaction concepts based on social context (namely culture and physical space). The complete validation of this work in the field of social-psychology would require an intense and continuous observation process of the same space. However, this effort represents a complete new methodology for social annotation of datasets in the field of computer vision.

The annotation is subdivided into two levels: i) *low-level features*, related to human detection and tracking, where a bounding box enclosing a person was marked on each frame, such information is represented by trajectories. Re-identification was not considered and when a person is partially or fully occluded, the bounding box was not marked. Also, a full-oriented gaze-direction $[0°, 360°]$ was annotated over person's head; ii) *high-level semantics*, related to individual

[1] Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto - http://sigarra.up.pt/fpceup



(a)

(b)

Fig. 1: (a) Detected chessboard points for camera calibration; (b) Horizontal vanishing line (blue), ground plane's projection area (green), ground points (red) to calculate scale factors and reprojection errors, and objects of interest (purple).

and collective profiles, where a trajectory could be composed by more than one profile. Also, objects of interest in the scene were marked, namely candy box, toy cars, and electric stairs (see Figure 1(b)). In terms of collective profiles, there are: i) *equally interested*, when a group presents a coherent behavior related to the same object (e.g. *looking to a storefront shop*); ii) *unbalance interested*, when a group reveals different types of of behavior in the scene at the same time; iii) *follow me*, when within a group a leader appears and restricts further group's behavior (e.g. *call the others to go by the stairs*). The individual profiles identified are: i) *distracted*, when no specific interest is revealed, translated into unstructured movement and high variability on gaze (e.g. *talking to a cellphone*); ii) *exploring*, when no specific interest is revealed, but movement and gaze are coherent with the scene structure and context (e.g. *looking for a pair of jeans*); iii) *interested*, when an interest by an object on the scene is explicitly revealed (e.g. *get a shopping car*). In this paper, we just analyze the individual profiles. Some of the most relevant statistics about the annotation are:

- annotated frames: 35866 (43$\%$ of the complete sequence);
- annotated persons: 98 (22 children, 40 men, 36 women);
- average frame count/elapsed time per individual profile: 276 ($\min$: 21, $\max$: 4260, σ: 445) / 11 sec;
- average frame count/elapsed time per group profile: 473 / 19 sec;
- average individual profiles per (person's) trajectory: 1.3.
- 130 individual profiles (26 *distracted*, 80 *exploring*, and 24 *interested*).

Since we are dealing with position and attention-based features, the trajectories should be projected onto the ground plane to correctly estimate distances and angles of interest. Such transformation involves camera calibration and geometry reconstruction steps that will be described in next Section.

IV. Camera Calibration and Ground-plane Projection

This stage assumes a relevant role to correctly achieve directional and geometry information to feed the proposed descriptor. We first acquire the camera parameters through calibration, and then estimate the vanishing lines over the rectified image. Both steps were combined sequentially to obtain ground-plane projection.

We did not have any physical scene measurement and camera knowledge at prior. However, we took advantage of a chessboard that appears on some frames, as showed in Figure 1(a), to get the image and objects points to proceed with the camera calibration. We just used the outer corners points since the inner color rectangles are not equal and are not aligned. We marked out the initial and final frames from which the chessboard is visible and we run an automatic technique to detect them. It is based on a simple template matching with rotation and scale invariance. In order to get relevant samples and better calibration results, we just took the frames where the points suffer a significant translation or rotation. Considering the physical proportions when compared with the person that holds the chessboard, we set its physical dimensions to ($0.7m$, $0.58m$). From this calibration pattern, we extracted the camera intrinsic and extrinsic parameters, as well as the undistorted and rectification matrices to apply on video frames.

To compute the vanishing points, ($\mathbf{v}_x$, $\mathbf{v}_y$, $\mathbf{v}_z$), we adopted the M-estimator SAmpling and Consensus (MSAC) algorithm [18]. Its input is a set of line segments that was automatically detected using a combination of the probabilistic Hough transform (PPHT) [19] with the Line Segment detection using Weighted Mean-Shift (LSWMS) [20]. For the former one, we used several incremental threshold levels, and set up a minimum line length and a maximum allowed gap between points on the same segment. These parameters reduce the erroneous lines normally detected on noisy regions. For the latter, we turned off the weighted Mean-Shift step, since it maintains accuracy and increase computational speed, and ranked out the line segments by orientation error and kept the K higher. We defined a maximum number of segments, S, to detect. Since, normally, the results of LSWMS are better than the PPHT, we account with more segments generated from LSWMS, where $K = \omega S$ (empirically $\omega = 0.8$). For the reminder, we just considered the results from PPHT that correspond to spatial regions where no valuable information from LSWMS was obtained. This combination permits to obtain reliable and complement line segments spatially distributed on regions with edges information.

Following the work of [21], we identified the vertical vanishing point, $\mathbf{v}_z$, and computed the horizontal vanishing line and the vanishing lines for the other planes in the scene. These geometric cues permit to get the direction of lines and the orientation of planes, and an image-to-world mapping for each image point could be calculated, the so-called homography matrix H. Indeed, considering the ground plane, $z = 0$, a plane-to-plane mapping can be defined, $X^T = H^{-1}x$, where x is the image point, X is the ground point and $H = [a\mathbf{v}_x \quad b\mathbf{v}_y \quad \mathbf{l}]$, where a and b are scale factors, and $\mathbf{l}$ is the horizontal vanishing line normalized.

The cross-ratio invariance concept [21] was used to compute the scale factors as well as the scale heigh factor. For both, physical measures should be known. In our case, we estimate them by physical relationships. For the height factor, we considered a linear trajectory that pass through the most farthest ground plane segment to the closest one, to account with perspective distortion, and considered a mean human height ($\approx 1.75m$). For the ground plane scale factors, we used the rectangle points identified on Figure 1(b) and approximated their measures to $(0.55m, 0.4m)$. For each point, their two components are calculated on the ground plane coordinate system defined by $\mathbf{v}_x o \mathbf{v}_y$. Since these points are aligned, we explored their geometric pattern to compute reprojection errors. To obtain gaze angle in the ground plane, both gaze direction vector's points were projected and angle was measure considering the defined ground plane coordinate system.

To obtain the ground plane's projection area, the user should indicate three points: the origin, a colinear point, where both define a segment, and a point located at the parallel and opposite segment. Using these points and the geometric information previously discussed, the projection area on the ground plane is automatically determined (see Figure 1(b)).

V. Semantic-context Descriptor

In this work, the descriptor is build over the annotated data. As mentioned on Section III, the annotation process imitates expected behavior of automatic trackers during occlusion situations. The objects of interest were also defined. Trajectories' characteristics extracted from our previous work [22] could give us those regions of interest, but we did not yet merge both works. Under these conditions, we can test the discriminative power of the descriptor, as well as analyze its features importance under the classification framework.

Our descriptor is inspired by [23]. Context-based features are incorporated into a key-point trajectory representation and transformed into a fixed-length descriptor to be used in a Bag-of-Features (BoF) approach. Each feature is encoded into a multi-scale histogram controlled by R, the number of granularity levels where the number of bins are given by 2^R, and the final descriptor is the concatenation of each feature's histogram. In this work, we tested a smoothing filter to get information from different timescales, but we leave it out since no significant difference was perceived among the various levels and further classification results corroborate such condition. We explain such behavior by the low spatial complexity and noise associated with the annotated data.

Measurements of relational position-based and attention-based cues are represented by key-points trajectory and captured by the descriptor, namely:

- *relative direction of movement*, α_{si}, is the orientation difference between subsequent trajectory segments. Final segment is considered equal to the previous one.
- *distance of interest*, d_{io}, expresses the distance between individual position and the object of interest.
- *direction of interest*, β_{gi}, which is the gaze direction.
- *velocity*, v_i, expresses the trajectory segment velocity.

Under our experiments, $R = 3$ showed a good trade-off between accuracy and dimensionality length, which leads to a 112-dimensional descriptor. Further classification framework details, as well as analysis of each feature's importance on final descriptor are presented on the following Section.

VI. Classification

The descriptor is fixed-length to be embedded into a BoF classification approach. In order to build the codebook, part of the annotated data was used to apply a k-means method over it. The obtained clusters form the vocabulary to be used on further training and classification processes.

We trained a binary classifier for each individual profile. Each trajectory's profile was subdivided temporally by a gap τ, which was treated as a bag. Each bag is considered a sample in the learning process. From the annotated data, we chose 50% of the profile's samples to be used for training. The number of positive and negative samples were the same and selected as the $\min$ of both, where the number of negatives is given by $\sum_{i<n, i\neq j} n_i$, n is the number of profile's samples and j the positive profile label. The final descriptor vector for each profile is a histogram obtained by nearest cluster counting, which was used as input for the linear SVM classifier. We adopted a k-fold cross-validation process to obtain the final classifier for each profile.

Since we are dealing with annotated data, we did not weight features to avoid noisy background features as explained in [23]. However, we did a backward feature selection to inspect their importance on final descriptor. We used two methods, namely gain information and relief-F. Since the descriptor obtained from the BoF approach is a histogram, we did a backward procedure starting from the discrete parts of the descriptor (clusters), until the individual feature bins. 1) *cluster ranking*, C_{r_i}, where to each cluster was applied a feature selection technique and an importance ranking was obtained; 2) *feature bin ranking*, F_{r_j}, on each cluster the previous step was applied again, resulting on a ranking of bins. Each bin corresponds to an individual feature, described on Section V. The final individual feature importance was obtained by

$$F_k = \sum_{i=0}^{C} \sum_{\substack{j=0,\\ l_j=k}}^{B} C_{r_i} \cdot F_{r_j}, \tag{1}$$

where C is the number of clusters, B is the number of bins on each cluster, and the condition $l_j = k$ permits to account only with feature bins that correspond to feature's label k.

VII. Results

Under this section we present the results relative to the most important parts of the work, namely camera reprojection and image-to-ground plane projection errors, features importance analysis and classification results. We evaluate these topics over two video sequences belonging to the extended IIT dataset (Section III) and the CAVIAR dataset[2], in specific the INRIA 1st set. For the latter, we adapted the *browsing* scenarios to our semantics. We considered that the *waiting* action performs a *distracted* profile. At the end, we account with 31 profiles, from which there are 4 *distracted*, 22 *exploring*, and 5 *interested*. Two objects of interest were marked, the brochure stand and the LCD screen stand.

During camera calibration, our automatic procedure permits to acquire image points with significant variations in pose (rotation) and depth (translation). This improves the fitting of the camera model, since bigger set of parameters, namely angles and positions, are given instead of redundant data that probably add noise. The final reprojection error is expressed as the average of the root mean square of the difference between the image points used to compute the object points and the recalculated image points from the obtained camera model. For the IIT video sequence the reported error is very low,

[2] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

0.117. We used the ground plane points, identified in Figure 1(b), to estimate the image-to-plane projection error. Taking in consideration the rectangle pattern between subsequent pair of segments, both in parallel and perpendicular directions, we computed the absolute errors of collinear segments in real world measure. The results vary between $[0.335\%, 4.908\%]$. For CAVIAR, we applied a simplified homography matrix, since the four image points that define the ground plane's projection area were given, as well as their physical coordinates.

To analyze our descriptor performance, we compare the classification results with a baseline descriptor, which is composed by the same type of features enumerated on Section V, but instead of considering a multiscale histogram based on key-point trajectories, it simply considers the mean, μ, and standard deviation, σ, of each feature. For the k-fold cross-validation, we used $k = 2$ and repeat it for 10000 iterations. The evaluation was done based on three standard parameters: accuracy (A), recall (R), and precision (P). Averaged results are reported on Table I. It is important to mention that for the CAVIAR sequence, we did not considered the gaze feature since it was not provided by the dataset.

TABLE I: Classification results on both video sequences.

IIT	Interested			Exploring			Distracted		
	A	R	P	A	R	P	A	R	P
Our	**88.0**	**93.8**	85.6	**85.7**	**94.0**	**81.4**	**89.0**	**80.9**	**96.9**
Base	86.9	82.7	**91.4**	68.9	88.4	64.2	76.5	79.0	76.7

CAVIAR	Interested			Exploring			Distracted		
	A	R	P	A	R	P	A	R	P
Our	91.9	84.6	**99.5**	78.2	83.7	**79.7**	**78.6**	**87.0**	80.5
Base	**92.6**	**94.9**	93.5	**81.4**	**94.4**	75.6	75.6	65.5	**88.4**

Considering the ITT sequence, our descriptor presents the highest results and the differences are significative. The baseline just presents a slightly better performance on the precision value of the *interested* profile, probably because it obtains lower false positive (FP) rate. However, inspecting Figure 2, we state that our descriptor produces low false negative (FN) and FP rates. Despite such low values, we analyze the distribution of FP entries per profile (Figure 3) and verify that the *exploring* and *interested* profile are interchangeable confuse. Specially, the *exploring* profile is the larger entry for the remaining profiles, probably because it is the one that has more samples.

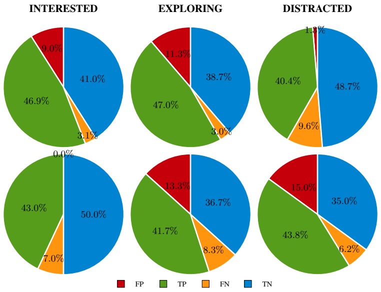


Fig. 2: Visualization of confusion matrix entries per profile (1st row: IIT; 2nd row: CAVIAR).

For the CAVIAR sequence, our descriptor presents less dominant results. Such behavior could be explained because we do not account with the gaze feature on this sequence, which has a relevant role (see Figure 4). This is also a prove that the multi-scale representation introduces a high discriminative factor. We also believe that another problem on this dataset was to translate the *waiting* action into a *distracted* profile. Intuitively, the *distracted* profile, along with the *exploring*, should present high variability in movements and gaze, which is clearly not the case for the *waiting* action. This is stated by a large FP rate for both profiles (*interested* profile has a FP rate close to 0, therefore none distribution is shown for this profile). In general our descriptor presents smaller standard deviation values of the three evaluation metrics than the baseline, for instance for IIT $[0.039, 0.104]$ against $[0.054, 0.168]$ and for CAVIAR $[0.040, 0.262]$ vs $[0.076, 0.316]$, which states that our model is more robust and stable than the baseline.

The feature selection process was conducted over the IIT sequence. As illustrated in Figure 4, both methods present the same importance hierarchy of features, which is also maintained along the profiles. It clearly shows that the relational context-based features are the most important, with relevance for the gaze. This corroborates our previous conclusion about the results obtained for the CAVIAR sequence that worst results are obtained when gaze feature is not considered. We preferred the relief-F method since it represents better the importance differences among features.

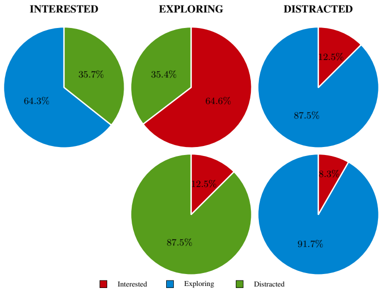


Fig. 3: Distribution of FP entries per profile (1st row: IIT; 2nd row: CAVIAR).

VIII. Conclusions

We presented a relational descriptor that considers contextual-based features to identify individual profiles. Such profiles were created upon a new semantic concept that follows an adjective association in a grammar-based analogy. We believe that they could help to form richer visual phrases for activity recognition. A real-life database, with intense and diverse human activity, was extended through manual annotation as a first effort to show the potential of the introduced semantic concept, as well as the proposed descriptor.

Our evaluation methodology confirms high rates of recognition. Comparison with a competitive baseline descriptor validates the relevance of the multi-scale histogram representation based on key-point trajectory sampling. We also presented a feature selection scheme that permits to inspect feature importance in a BoF-based descriptor. Its intuition is novel and validates the pertinence of the relational context-based features. The use of the CAVIAR scenario shows the scalability of the profiles into the existing concepts of human activity. However, more sequences should be tested to deeper inspect

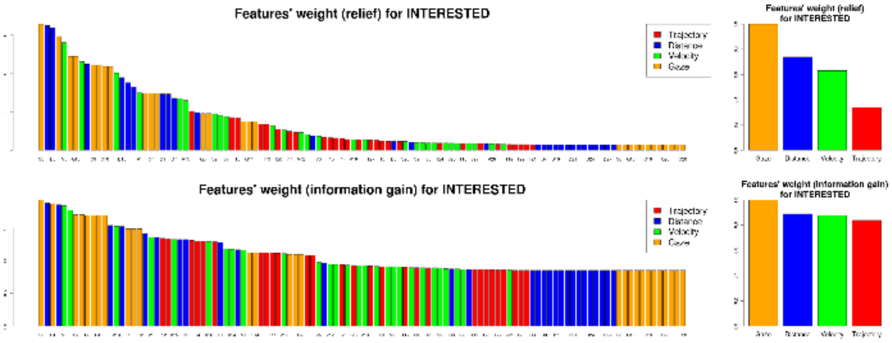


Fig. 4: Features importance analysis: trajectory orientation, trajectory velocity, gaze and distance to object of interest.

profile's impact and discriminative descriptor power. Future work will follow two directions: extension of the descriptor to collective behavior, and embedding of the descriptor into an unsupervised learning approach.

Acknowledgment

This work is financed by National Funds through the FCT - Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within PhD grant reference SFRH/BD/51430/2011, and post-doctoral grant SFRH/BPD/85225/2012. The authors would like to thank Amit Adam for supplying the video sequences, Kelly Rodrigues and the Social Psychology Research Group of the University of Porto for their scientific advice.

References

[1] G. Pusiol, F. Bremond, and M. Thonnat, "Trajectory Based Activity Discovery," in *7th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Boston, États-Unis, Aug. 2010. [Online]. Available: http://hal.inria.fr/inria-00504634

[2] W. Ge, R. T. Collins, and B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, 2012.

[3] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 3, pp. 397–408, 2005.

[4] J. Owens and A. Hunter, "Application of the self-organizing map to trajectory classification," in *Proceedings of the Third IEEE International Workshop on Visual Surveillance (VS'2000)*, ser. VS '00. Washington, DC, USA: IEEE Computer Society, 2000, pp. 77–. [Online]. Available: http://dl.acm.org/citation.cfm?id=832293.836186

[5] S. Khalid and A. Naftel, "Classifying spatiotemporal object trajectories using unsupervised learning of basis function coefficients," in *VSSN '05: Proceedings of the third ACM international workshop on Video surveillance &amp; sensor networks*. New York, NY, USA: ACM, 2005, pp. 45–52. [Online]. Available: http://dx.doi.org/10.1145/1099396.1099404

[6] X. Wang, K. Tieu, and E. Grimson, "Learning semantic scene models by trajectory analysis." pp. 110–123, 2006. [Online]. Available: http://dblp.uni-trier.de/db/conf/eccv/eccv2006-3.html#WangTG06

[7] X. Wang, K. T. Ma, G.-W. Ng, and W. E. Grimson, "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models," *Int. J. Comput. Vision*, vol. 95, no. 3, pp. 287–312, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1007/s11263-011-0459-6

[8] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, no. 5, pp. 4282–4286, May.

[9] F. Klgl and G. Rindsfser, "Large-scale agent-based pedestrian simulation." in *MATES*, ser. Lecture Notes in Computer Science, P. Petta, J. P. Mller, M. Klusch, and M. P. Georgeff, Eds., vol. 4687. Springer, 2007, pp. 145–156. [Online]. Available: http://dblp.uni-trier.de/db/conf/mates/mates2007.html#KluglR07

[10] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *ECCV (2)*, ser. Lecture Notes in Computer Science, D. A. Forsyth, P. H. S. Torr, and A. Zisserman, Eds., vol. 5303. Springer, 2008, pp. 1–14.

[11] M.-C. Chang, N. Krahnstoever, and W. Ge, "Probabilistic group-level motion analysis and scenario recognition," in *ICCV*, 2011, pp. 747–754.

[12] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1922649.1922653

[13] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and F.-F. Li, "Human action recognition by learning bases of action attributes and parts," in *ICCV*, 2011, pp. 1331–1338.

[14] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities." in *ICCV*. IEEE, 2009, pp. 1593–1600. [Online]. Available: http://dblp.uni-trier.de/db/conf/iccv/iccv2009.html#RyooA09

[15] B. Jin, W. Hu, and H. Wang, "Human interaction recognition based on transformation of spatial semantics." *IEEE Signal Process. Lett.*, vol. 19, no. 3, pp. 139–142, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/spl/spl19.html#JinHW12

[16] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR*. IEEE, 2011, pp. 3273–3280.

[17] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors." *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 3, pp. 555–560.

[18] M. Nieto and L. Salgado, "Non-linear optimization for robust estimation of vanishing points," in *ICIP*. IEEE, 2010, pp. 1885–1888.

[19] J. Matas, C. Galambos, and J. Kittler, "Robust detection of lines using the progressive probabilistic hough transform." *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 119–137, 2000. [Online]. Available: http://dblp.uni-trier.de/db/journals/cviu/cviu78.html#MatasGK00

[20] M. Nieto, C. Cuevas, L. Salgado, and N. N. García, "Line segment detection using weighted mean shift procedures on a 2d slice sampling strategy," *Pattern Anal. Appl.*, vol. 14, no. 2, pp. 149–163, 2011.

[21] A. Criminisi, "Accurate visual metrology from single and multiple uncalibrated images," Ph.D. dissertation, University of Oxford, Dept. Engineering Science, 1999, d.Phil. thesis.

[22] E. M. Pereira, J. S. Cardoso, and R. Morla, "Motion flow tracking in unconstrained videos for retail scenario," in *Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science, J. a. Sanches, L. Micó, and J. S. Cardoso, Eds., vol. 7887. Springer Berlin Heidelberg, 2013, pp. 340–349.

[23] M. Takahashi, M. Naemura, M. Fujii, and S. Satoh, "Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories," in *Computer Vision and Pattern Recognition*, 2011, pp. 9–16.