Motion Flow Trajectories: The Beginning of a Visual Perceptual Reasoning System in a Unified Data-Driven Approach

Eduardo Marques
ejmp@inescporto.pt
Jaime S. Cardoso
jaime.cardoso@inescporto.pt
Ricardo Morla
ricardo.morla@fe.up.pt

INESC TEC and Faculdade de Engenharia
Universidade do Porto, Porto, Portugal

Abstract

This paper has a two-fold intention: i) to describe our global research proposal; ii) to present our on-going work towards the understanding of semantic human behavior in a retail complex scenario, which ultimate goal is to build a behavioral semantic model that analyzes shopper's non-verbal features individually and in group, their relations with the scene, and their needs for assistance support.

1 Introduction

Human activity understanding is an important and growing area of computer vision research.It is an area that explores automatic reasoning mechanisms to express human activities by high-level semantics, which can be composed for multiple atomic actions, entities relationships, and context information. However, until the recent years, this area was predominant influenced by recognition methodologies that just analyze human body taxonomies. Thanks to psychological, sociological and cognitive studies, increasing of computational hardware power, and to the advance of parallel areas in computer vision such as object recognition, scene layout recovery, and scene interpretation, most recent works are starting to merge mutual context with the aim to improve human activity understanding.

At the level of human-human interactions, Jin *et al*. [7] used context-free grammar to recognize interactions, based on production rules governed by semantic spatial relationships between individuals obtained from multi-tracking algorithms. Choi *et al*. [2] recognized collective human activities by modeling common and repetitive configuration behaviors of individuals, and their relationships distributions over time and space.

Considering relations between human-space, long-term observations of moving objects in the scene allow to build semantic scene models from the spatial distribution of trajectories. Pusiol *et al*. [10] designed an intermediate layer composed of Primitive Events descriptors, to cluster motion trajectories by individual segments of slow points in trajectories and extract meaningful transition between topological slow regions.

The work of Gupta *et al*. [4] explored relations between scene and objects within it, by building a physical representation that models objects as volumetric entities, and their relationships to describe the 3D structure and mechanical layout of the scene.

This section highlights an important change of research paradigm that states that the understanding of 3D activity should be treated as one complex problem since the beginning, rather than divided by a number of independent detached problems. Next, we present an overview of our research proposal in section 2, followed by a preliminary work on a real application scenario, section 3. Finally, we formulate our conclusions and directions for future work in section 4.

2 Global research proposal

Our proposal refers to an unsupervised perceptual framework that can help us to understand how to identify and classify high-level human activity in indoor environments at multiple set of temporal scales under an specific application context. We are interested on integrating microscale representation with context modeling to infer characterization of human activity.

Our statement considers: i) ***Active-Learning of Concept-Characterizations:*** [5] proved that activity classes could be modeled as combinations of their local sequential features. Our aim is to add perceptual information about the elements of activity dynamics at any stage to improve inference of concepts and categories; ii) ***Object and Movement Space Mapping:*** action primitives [8] are used for both recognition and synthesis. We believe that their associated grammatical description could be integrated as activity descriptors; iii) ***Discriminative Context Modeling:*** [11] showed that incorporation of shape, appearance and context is an efficient method for image categorization and segmentation. We believe that this approach could benefit from semantic and spatial relations; iv) ***3D Spatial Structure:*** [12] proved that spatial structure reflects the functionality of the location and help to suggest scene category. We believe in a holistic approach for recognize functional scene layout.

Our novelty resides in an unified approach that deals with the interpretation of objects, scene, actions, and their mutual contextual constraints to improve action classification, scene context categorization, and semantic inferring. We believe that application goals will largely benefit from this perceptual framework, and as an example, we present next, in section 3, a brief description of our first approach, based on motion characteristics, to extract human behaviors on shopping scenario.

3 Human behavior in retail scenario

Understanding customers behaviors and their purchase decision processes in an automatic way is an inestimable commercial advantage for the retail market. These kind of applications have been facing a growing demand; However, there is not any fully automated system for customer behavior analysis commercially available.

Our major contribution is in the comparative performance of two systems to compute global motion field for path learning in a very challenging scenario: a video with small resolution, low frame rate (1 fps), and uncontrolled camera deployment process of a real cluttering shopping store with heavy crowds, occluded bodies, and reflective surfaces, where customers move, individually and in group, randomly in various directions, and whose appearance models changes abruptly from frame to frame.

The traditional approach for motion analysis consists of detecting objects, tracking them, and analyze their tracks for event/activity detection. This standard processing does not work well on high density scenes with cluttered environment since reliable trajectories of objects can not be obtained. To solve these cases, global motion flow field is used to learn typical motion patterns. Under these constraints and considering motion field acquired from optical flow algorithms, we implemented and evaluated two approaches: grid-based global dominant motion flow method, and kernel-based sink-seeking method. These methods belong to the last processing step of the baseline framework composed by the following steps: sampling, motion flow fields, matching, and motion flow tracking.

The sampling process is responsible to extract a set of relevant points for frame matching, and can be subdivided into two types: i) dense, ii) sparse. Normally, for image classification and action recognition dense sampling performs better than sparse sampling, since it enables local reasoning from motion similarities, and introduces spatial regularity constraints in the clustering method. However, we verified empirically that for our dataset sparse sampling is preferred, not only for computation effort when generating the global motion trajectories, but also because dense sampling introduces noisy points from the cluttered background, and does not increase discriminative value for trajectories.

The motion flow fields are obtained from existing optical flow algorithms that consider frame-to-frame analysis or larger spatio-temporal displacements. This step is computed independently, and the resulting motion flow map is used in the matching step process, in conjunction with the sampling points. Since our dataset has very hard demanding conditions, we considered two types of optical flow algorithms for motion estimation: i) short-classical, ii) large displacement. We verified that this step

largely influences the final output. From several algorithms tested, we highlight two of them: the short-classical Farnebäck [3], and the descriptor matching in variational model (LDOF) [1]. Both of them present good results, however the latter is an offline method that permits the extraction of smoother and longer trajectories, and the former is computed on real time but introduces noise on areas with large appearance variations.

The matching step consist in the tracking of each sampling point $P_t = (x_t, y_t)$ from frame t to next frame $t+1$, to obtain point $P_{t+1} = (x_{t+1}, y_{t+1})$. These two points form a flow vector at a specific pixel location. The adopted approximation is a median filtering kernel, that performs better than the bilinear interpolation and removes impulse noise, preserves edges, and smooths points in dense optical flow fields.

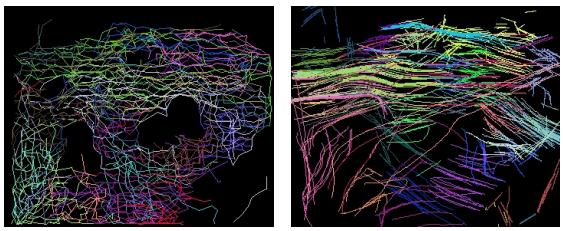
The motion flow tracking step has a common pipeline to extract meaningful flow vectors. The image region is divided by an equally spaced grid, and each cell contains the flow vectors that lay inside it. On each cell is applied a two-step hierarchical clustering approach with a two-fold purpose: to reduce the number of flow vectors, still maintaining the geometric structure of the flow field, and to obtain the local dominant motion flows. The first clustering step considers a full-orientation histogram with 8 bins to express the orientation groups. The groups with weight above the histogram's mean are taken into account for next clustering step. The second step implies a spatial clustering on each orientation group. At the end, there are several clusters for each orientation group that are ordered in a descendent-weighted way, considering the number of flow vectors that belong to them. These groups represent the local dominant flows.

3.1 Results

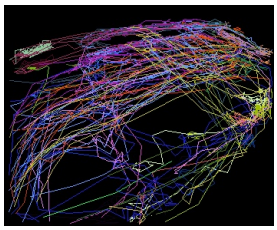
The two implemented algorithms of the final processing step are: i) ***Grid-based Global Motion Flow:*** computes the global dominant motion flows derived from local dominant motion flows in neighborhood, and it is based on Ozturk*et al*. [9] approach. Each cell has a depth-step for looking for neighbors, which are defined as the regions that are in the direction of the current local flow. The scanning process considers a first iteraction, where the current dominant flow just search for neighbors that belong to the same orientation group, and in case of no returning the subsequent iteraction consider the closest flow in the valid neighborhood to connect with it, which are the adjacent orientation groups. If no neighbor flow is obtained in the valid depth-step, any orientation is considered to keep continuity and permit abrupt flow orientation changes; ii) ***Kernel-based Sink-Seeking:*** follows the work of Hu*et al*. [6], and considers a sink-seeking process, which is a sliding-window-based technique that uses a continuous kernel-based estimator to obtain global motion paths. It incorporates neighborhood flow motion information to obtain the representative states of the global motion path. The relation between next and previous states is linear, and the next location depends on position and flow motion of neighboring points. The neighborhood is defined by the flow vectors that lay inside the kernel window, and whose angle difference against the current angle sink state is below a certain threshold. The linking between consecutive sink-seeking states gives the global motion flow.

Both algorithms extract meaningful motion trajectories indicating the global motion flows in cluttered regions, however, they perform differently. A global comparison: the kernel-based method produces smoother but shorter trajectories than the grid-based, probably since the appearance changes a lot from frame to frame, the kernel window does not accept flow vectors with large opposite directions, so the sink-seeking process maintains a coherent flow, and since it is a sliding window technique, its overlap permits to build smoother trajectories; the kernel-based method is much slower, since the sink-seeking process is repeated for each flow vector; the kernel-based method produces much more trajectories than grid-based, so it requires a sparse sampling and just one local dominant flow vector by cell to run on real time; the grid-based method is more sensitive to noise and does not keep an smooth flow.

Individually each algorithm has a combinatorial number of factors that affects the final performance. The grid-based method's results are affected by: i) the *pair-wise metric* used to choose the next local dominant flow from the returned neighborhood's local flows, which can be just spatial, or a weighted additive distance; ii) the *neighborhood*, which has a depth-step factor that affects the search radius among cells, and the similarity metric between current and next orientation group, which can be equal, similar, or other, provoking the continuity and/or discontinuity of flow. In turns, the kernel-based method presents other factors: i) the *bandwidth size*, which largely affects the sink-seeking process and has three important steps that influence results: size initialization, size restart on each vector flow, and size updating; ii) the *acceptance angle*, which is a threshold to filter the flow vectors that participate on the kernel window, the greater more probabilities to have opposite vector flows to push current trajectory to other directions.



(a) Grid-based trajectory clusters. (b) Kernel-based trajectory clusters.



(c) Ground-Truth trajectory clusters.

Figure 1: Clustering results for $k = 39$, where k is the number of clusters.

Ground truth data was obtained from manual tracking annotation, an for each person two reference points were considered: head and center of mass. A parallel spectral clustering, with K-means on its final stage, was applied on distance matrices of trajectories from ground truth data, and from result data of both algorithms. The results are presented on figure 1.

4 Conclusions

In this paper we briefly describe our research proposal justified for its relevance on scientific community, and explain its purpose and foundations. A real case scenario for understanding shopping behavior is summarized and its qualitative results are discussed. This reflects a work-in-progress, whose next step will be the validation of exact metrics to evaluate the performance on computing trajectories based on motion flow, a trajectory encoding scheme to estimate space layout and semantic local regions, and implement a tracking social model to infer human-human relationships.

Acknowledgements

This work is financed by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701, and for the PhD grant with reference SFRH/BD/51430/2011.

References

[1] Thomas Brox and Jitendra Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):500–513, March 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.143. URL http://dx.doi.org/10.1109/TPAMI.2010.143.

[2] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR*, pages 3273–3280. IEEE, 2011.

[3] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian conference on Image analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag. ISBN 3-540-40601-8. URL http://dl.acm.org/citation.cfm?id=1763974.1764031.

[4] Abhinav Gupta, Alexei A. Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *European Conference on Computer Vision*, 2010.

[5] Raffay Hamid, Siddhartha Maddi, Amos Johnson, Aaron Bobick, Irfan Essa, and Charles Isbell. A novel sequence representation for unsupervised analysis of human activities. *Artif. Intell.*, 173(14):1221–1244, September 2009. ISSN 0004-3702. doi: 10.1016/j.artint.2009.05.002. URL http://dx.doi.org/10.1016/j.artint.2009.05.002.

[6] Min Hu, Saad Ali, and Mubarak Shah. Detecting global motion patterns in complex videos. In *ICPR*, pages 1–5. IEEE, 2008. ISBN 978-1-4244-2175-6.

[7] Biao Jin, Wenlong Hu, and Hongqi Wang. Human interaction recognition based on transformation of spatial semantics. *IEEE Signal Process. Lett.*, 19(3):139–142, 2012.

[8] Dana Kulic, Danica Kragic, and Volker Krüger. Learning action primitives. In Thomas B. Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal, editors, *Visual Analysis of Humans*, pages 333–353. Springer, 2011. ISBN 978-0-85729-996-3.

[9] Ovgu Ozturk, Toshihiko Yamasaki, and Kiyoharu Aizawa. Detecting dominant motion flows in unstructured/structured crowd scenes. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 3533–3536, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4109-9. doi: 10.1109/ICPR.2010.862. URL http://dx.doi.org/10.1109/ICPR.2010.862.

[10] Guido Pusiol, François Bremond, and Monique Thonnat. Trajectory Based Activity Discovery. In *7th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Boston, États-Unis, August 2010. URL http://hal.inria.fr/inria-00504634.

[11] Jamie Shotton, John M. Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[12] Jianxin Wu and James M. Rehg. Centrist: A visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1489–1501, 2011.