Estimating Human Actions Affinities Across Views

Nicoletta Noceti[1], Alessandra Sciutti[2], Francesco Rea[2], Francesca Odone[1] and Giulio Sandini[2]
[1]*DIBRIS, Università di Genova, via Dodecaneso 35, 16146, Genova, Italy*
[2]*RBCS, Istituto Italiano di Tecnologia, via Morego 30, 16163, Genova, Italy*

Keywords: Human Actions Understanding, View-invariant Representation, Trajectory Curvature.

Abstract: This paper deals with the problem of estimating the affinity level between different types of human actions observed from different viewpoints. We analyse simple repetitive upper body human actions with the goal of producing a view-invariant model from simple motion cues, that have been inspired by studies on the human perception. We adopt a simple descriptor that summarizes the evolution of spatio-temporal curvature of the trajectories, which we use for evaluating the similarity between actions pair on a multi-level matching. We experimentally verified the presence of semantic connections between actions across views, inferring a relations graph that shows such affinities.

1 INTRODUCTION

Since birth human neonates present a preference for biological motion (Simion et al., 2008) and this is an important trigger for social interaction. This inclination – a key element in human development research – is also inspiring computer vision researchers to finding computational models able to replicate it on artificial systems.
The contamination between the two research communities may also guide the choice of the most appropriate tools for a given computer vision task: if neonates have this capability of "understanding" motion while their visual perception is still rudimental, *it is likely* that the analysis is based on very simple motion cues – such as sensitivity to local apparent motion, and simple features as velocity, acceleration or curvature (see e.g. (Kuhlmeier et al., 2010; Simion et al., 2008)).
The goal of our research is to build computational models for the iCub humanoid robot (Metta et al., 2008) to simulate this phase of human development. Our long term objective is to understand how much we can infer on the nature of motion from such simple observations, as this skill seems to be at the basis of human ability in interacting with others. Also we would like to preserve some abilities typical of a developing human being, such as a degree of tolerance to view-point changes (Troje and Westhoff, 2006; Goren et al., 1975; Farah et al., 1995)

On a shorter term, our research is focusing on the identification of biological motion (some preliminary results can be found in (Sciutti et al., 2014)) and on the analysis of different types of biological motion. The latter is the goal of this paper, in which we analyze simple repetitive upper body human actions with the goal of producing a view-invariant model from simple motion cues. We apply this model to the estimate of the affinity level between different types of actions, focusing in particular on two main categories: *transitive* – which involve object manipulation – and *intransitive* actions.
Since we are primarily interested in capturing abilities typical of the early months of human development we do not address classical action recognition tasks, abilities which are likely to be gained in later stages of development (Camaioni, 2004; Kanakogi and Itakura, 2011), also thanks to the infants prior motor experience .

Our model takes inspiration from the seminal work (Rao et al., 2002), where the authors discuss on the use of dynamic instants, i.e. meaningful action units whose properties are highly characterizing the evolution of structured activities (e.g. *Pick up an object from the floor and put it on the desk*) and that have been proved to be of substantial relevance for the human motion perception. Such points – consequence of a variation in the force applied during an activity – correspond to the local maxima of the curvature of the trajectory describing the activity evolution on the image plane. The authors formally prove that they also have view-invariant properties.

In this paper we focus instead on *intervals*, mean-

Noceti N., Sciutti A., Rea F., Odone F. and Sandini G..
Estimating Human Actions Affinities Across Views.
DOI: 10.5220/0005307801300137
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 130-137
ISBN: 978-989-758-090-1

ing portions of trajectories between two dynamic instants, and investigate on their potentially informative content to be exploited for cross-view recognition. For our investigation, we collected a data set of videos, each one including repetitions of a given atomic action. Indeed, in our case dynamic instants mainly refer to points partitioning one action instance from the next one. After an initial low-level analysis, that aims at segmenting the moving arm of the user, we extract corner points from the arm region and track them over time to collect a set of trajectories. Then, we measure the curvature and find its local maxima. Finally, we describe the intervals with histograms of curvature, that we adopt in a multi-level matching to evaluate the level of affinity between two observed events. We experimentally provide the evidence of the presence of actions classes inferred by estimating the pairwise similarities of action sequences from the same view and across different views.

Works related to the computational model we consider can be found in fields as video surveillance, video retrieval and robotics, where tasks as *gesture and action recognition* or *behavior modeling* have been for many years very fertile disciplines, and still are (Fanello et al., 2013; Malgireddy et al., 2012; Mahbub et al., 2011; Noceti and Odone, 2012; Wang et al., 2009) We refer the interested reader to a recent survey (Aggarwal and Ryoo, 2011) for a complete account on the topic.

From the *view-invariance* standpoint, the problem has been addressed considering two different settings, i.e. observing the *same* dynamic event simultaneously from two (or more) cameras (Zheng et al., 2012; Wu and Jia, 2012; Li and Zickler, 2012; Huang et al., 2012a; Zheng and Jiang, 2013) or considering different instances of a same concept of dynamic event (Lewandowski et al., 2010; Gong and Medioni, 2011; Junejo et al., 2011; Li et al., 2012; Huang et al., 2012b). The latter are more related to our setting. In general, view-invariance may be addressed at a descriptor level (Junejo et al., 2011; Li et al., 2012; Huang et al., 2012b) or at the similarity estimate level. In this case machine learning (Wu and Jia, 2012; Huang et al., 2012a; Zheng and Jiang, 2013) and, more recently, transfer learning (Zheng et al., 2012; Li and Zickler, 2012) may be beneficial.

The approach we follow shares computational tools and models with many of the above mentioned works, but significantly differs in the intentions, in that we are not interested in recognizing specific gestures, actions or activities, but instead we consider a more abstract task: *to what extent are we able to infer properties (if any) on the observed motion that persist across views, even from a very coarse representation?*.

The rest of the paper is organized as follows. Sec. 2 is devoted to the description of the approach we follow, from the low-level analysis to the matching, while Sec. 3 describes our experimental analysis. The final section is left to conclusions.

2 CURVATURE-BASED MOTION REPRESENTATION

In this section we discuss our approach to motion modeling, that builds on top of a sparse representation and then relies on the computation of histograms of the spatio-temporal curvature. Then, we describe the strategy we adopt to match image sequences.

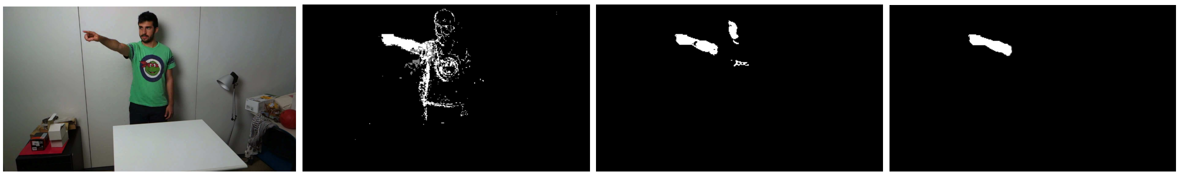
2.1 Low-level Video Analysis

The very first step of our method relies on a widely accepted video analysis pipeline, that aims at segmenting each image of a sequence with respect to both motion and appearance information. Instantiated to our case study, this corresponds to detecting the image region with the moving arm of a subject while performing a given action. The intermediate steps of the pipeline are reported in Fig. 1. We first perform background subtraction (Zivkovic, 2004), then refine the obtained binary map by applying skin detection in the HSV color space to only the foreground. Finally, assuming only the subject of interest is moving in the observed scene, we keep the largest connected component of the final binary map as region of interest (ROI).

The second stage of our method relies on describing the visual dynamic of the moving region by means of points trajectories (Fig. 2). To this purpose, we extract from the ROI the Harris corners (Shi and Tomasi, 1994), which we describe using SIFT descriptors (Lowe, 2004). Then, we track SIFTs with a Kalman filter (Welch and Bishop, 1995) and using histogram intersection as a similarity measure between observations. To improve the quality of the obtained spatio-temporal trajectories, we finally filter them with anisotropic diffusion (Perona and Malik, 1990). We collect spatio-temporal trajectories for each video, and thus set the basis for the next step of motion representation, based on the concept of curvature.

2.2 Spatio-temporal Curvature

The projection of the dynamic evolution of a 3D point in the image plane is composed as a spatio-



(a) (b) (c) (d)

Figure 1: A visual sketch of the initial video analysis: starting from an image of the sequence (Fig. 1(a)), we first apply background subtraction (Fig.1(b)), then detect skin on the foreground (Fig. 1(c)), and finally keep only the largest connected component (Fig. 1(d)) as the region of interest.

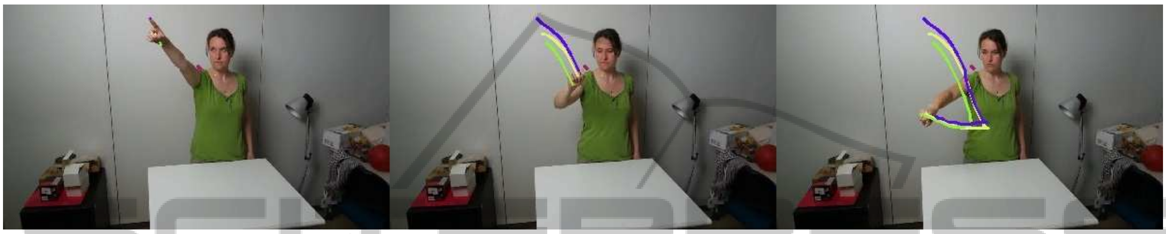


Figure 2: Examples of trajectories of corner points, which we subsampled for the sake of clarity of the figure.

temporal trajectory of observations $T = \{P(t_i)\}_{i=1}^{n}$, where $P(t_i) = (x(t_i), y(t_i), t_i)$ is an image point observed at time t_i with coordinates $(x(t_i), y(t_i))$, which are functions of time. The velocity of the points can be expressed as the derivative of the positions, i.e. $V(t_i) = (x'(t_i), y'(t_i), \Delta_t)$ where Δ_t is the temporal gap between adjacent images. Similarly, the acceleration is the derivative of the velocity, $A(t_i) = (x''(t_i), y''(t_i), 0)$. At this point, we can compute the trajectory curvature as

$$C(t_i) = \frac{||V(t_i) \times A(t_i)||}{||V(t_i)||^3}. \qquad (1)$$

Consider the corner trajectories of Fig. 2, one of which is reported in the 3D space-time reference system of Fig. 3, left. On the right, we show the trend of velocity magnitude (above) and of the curvature (below) over time, enhancing their local maxima in red and green, respectively. The corresponding space-time points are coherently marked on the 3D visualization on the left. As it can be easily observed, while the first type of points indicates time instants in the middle of a segment of the space-time trajectories (points in which the user starts to decelerate to finally stop the movement), the latter refers to instantaneous changes in motion direction and/or dynamic.

The dynamic instants are the units on top of which we build the representation of a trajectory. Following the notation of the previous section, an observed spatio-temporal sequence $T = \{(x(t_i), y(t_i), t_i)\}_{i=1}^{n}$ is now associated with a sequence $DI = [\hat{t}_1 \ldots \hat{t}_m]$ of dynamic instants, where $\hat{t}_i \in \{t_1, \ldots, t_n\}$ and $m < n$.

According to (Rao et al., 2002), we define an interval as a trajectory segment laying in the middle of two dynamic instants. We chose to represent the distribution of the curvature in it by means of histograms. Therefore, at the end of the representation, each observed trajectory is associated with a sequence of curvature histograms, i.e. $H(T) = [H(\hat{t}_1, \hat{t}_2), \ldots, H(\hat{t}_{m-1}, \hat{t}_m)]$, where $H(\hat{t}_i, \hat{t}_{i+1})$ refers to the fact that the histograms are computed between each pair of adjacent dynamic instants.

2.3 Multi-level Matching between Image Sequences

Once we have detected the dynamic instants as the local extrema of the curvature and represented the intervals curvature as histograms, we may set up a multi-level procedure to match image sequences. Hence, let us consider two videos and the two corresponding sets of observed trajectories, $\mathcal{T}^1 = \{T_1^1, \ldots, T_N^1\}$ and $\mathcal{T}^2 = \{T_1^2, \ldots, T_M^2\}$, described with the curvature histograms to collect the sets $\mathcal{H}^1 = \{H(T_1^1), \ldots, H(T_N^1)\}$ and $\mathcal{H}^2 = \{H(T_1^2), \ldots, H(T_M^2)\}$.

To match two image sequences, we start by comparing pairs of trajectories of type (T_i^1, T_j^2), with $1 \leq i \leq N$ and $1 \leq j \leq M$. For the sake of clarity, we express the sequence of histograms with a more compact style as $H(T_i^1) = [H_{i,1}^1, H_{i,2}^1, \ldots, H_{i,m_i-1}^1]$ and $H(T_j^2) = [H_{j,1}^2, H_{j,2}^2, \ldots, H_{j,m_j-1}^2$. Since the videos we consider refer to repetitions of a given atomic actions, we consider the average similarity between all pairs of histograms describing portions of the two trajectories. The similarity between two trajectories is thus

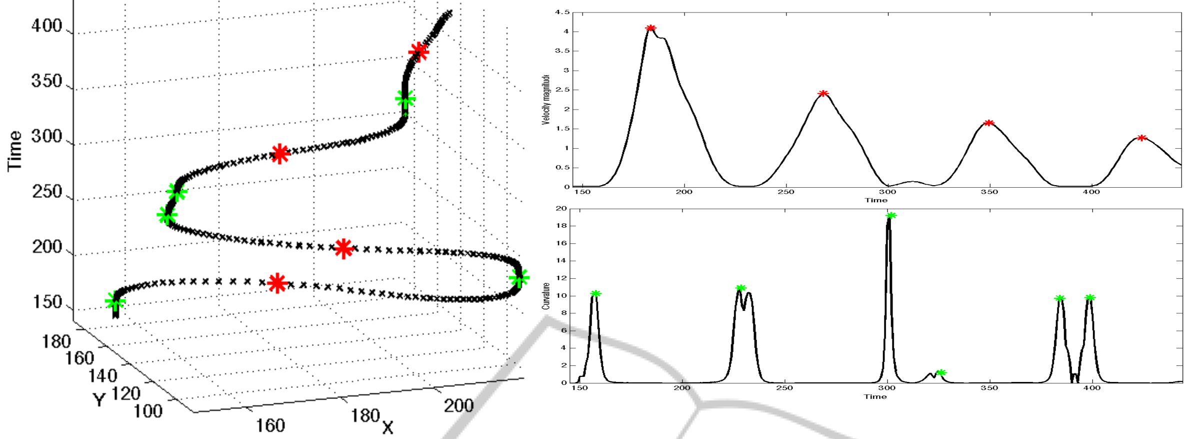


Figure 3: Left: a space-time representation of a trajectory. Right: velocity magnitude (above) and curvature (below) as functions of time. We denote local maxima of the velocity magnitude (red) and of the curvature (green). The latter correspond to dynamic instants, relevant for the human perception of motion.

formalized as

$$S(T_i^1,T_j^2)=\frac{1}{(m_i-1)(m_j-1)}\sum_{k=1}^{m_i-1}\sum_{h=1}^{m_j-1} HI(H^1_{i,k},H^2_{j,h}) \qquad (2)$$

where HI denotes the intersection between histograms. Given a video, the observed trajectories describe the evolution of 3D points all related to the same physical event, i.e. the motion of the user arm. Thus, it is convenient to summarize the contribution of all the trajectories to end up with a single value quantifying the global similarity between two videos, i.e. two physical events. To this purpose, we select the maximum similarities between all pairs of trajectories. More formally

$$S(\mathcal{T}^1,\mathcal{T}^2)=\max_{i=1\ldots N,j=1\ldots M} S(T_i^1,T_j^2). \qquad (3)$$

3 EXPERIMENTAL ANALYSIS

In the following we report on the experiments we performed in order to evaluate the level of view-invariant information included in the intervals between dynamic instants, which we describe and compare according to the previous section. Our main objective is to extract knowledge about properties of (classes of) actions that might be captured across views with this somehow primitive representation. To this end we performed a qualitative evaluation on a dataset we collected in-house.

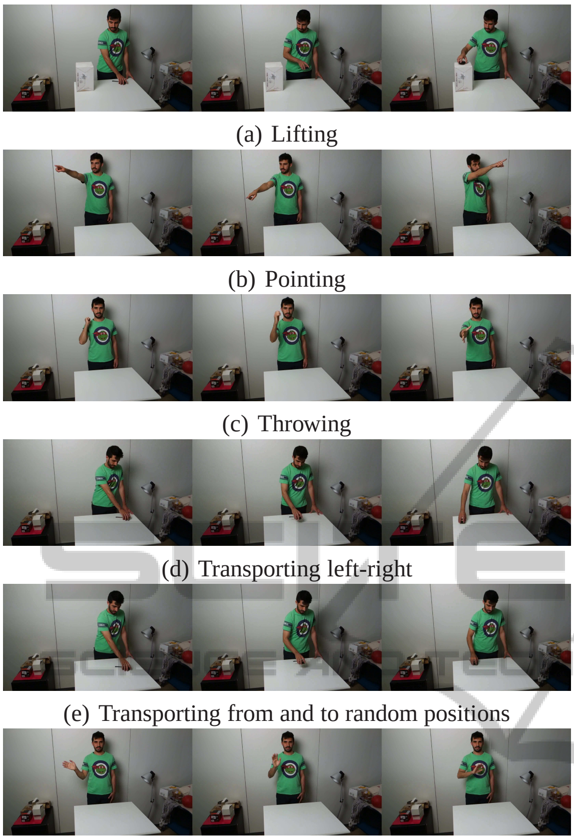
3.1 Experimental Setup

We acquired a set of image sequences of two subjects observed from two different viewpoints. The acquisitions have been made on an indoor environment to favor the low-level analysis and thus allow a higher focus on the second step of motion representation and matching. The variation of the viewpoint reflects the application we have in mind, i.e. human robot interaction, where we can assume the interacting subject to be located in a limited radial spatial range in front of the camera (i.e. the robot). Similarly, the actions included in the data set (shown in Fig. 4) are suggested by the application. We considered 6 actions: *Pointing* a finger towards a certain 3D location; *Waving* the hand from left to right and vice-versa; *Lifting* and object from the table to a box placed on it; *Throwing* an object away (action that we only simulated for practical reasons); *Transporting* an object from and to different positions on the table. The latter is instantiated in two versions, with left-right and random object repositioning.

Each video consists of 20 repetitions of the same atomic action (e.g. move the object from left to right); for each subject we acquired two videos in each view for each action.

3.2 Proof of Concepts

Spatio-temporal Curvature. After having extracted corners trajectories from each video, we first detect the dynamic instants. Let us start our analysis by providing an experimental evidence of the information carried by the dynamic instants in the setting we consider. In Fig. 5 we show examples



(a) Lifting

(b) Pointing

(c) Throwing

(d) Transporting left-right

(e) Transporting from and to random positions

(f) Waving

Figure 4: Samples from the acquisitions of a subject from a single viewpoint.

Table 1: Ranking from the comparison of videos of the same subject and acquired from the same viewpoint.

Test	Ranking					
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
Lift	**Li**	T.LR	T.Rnd	Wa	Po	Th
Point	T.Rnd	**Po**	Th	Li	T.LR	Wa
Throw	**Th**	Po	T.Rnd	T.LR	Li	Wa
T. LR	**T.LR**	Li	T.Rnd	Wa	Th	Po
T. Rnd	**T.Rnd**	Po	T.LR	Li	Wa	Th
Wav	**Wa**	T.Rnd	T.LR	Li	Po	Th

Table 2: Ranking from the comparison of videos of different subjects, acquired from the same viewpoint.

Test	Ranking					
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
Lift.	T.Rnd	T.LR	**Li**	Th	Po	Wa
Point.	Th	**Po**	T.LR	Li	T.Rnd	Wa
Throw.	Po	**Th**	Li	T.Rnd	T.LR	Wa
T. LR	T.Rnd	**T.LR**	Li	Po	Th	Wa
T. Rnd	Li	**T.Rnd**	T.LR	Th	Po	Wa
Wav.	**Wa**	Th	Po	T.LR	T.Rnd	Li

Table 3: Ranking from the comparison of videos acquired acquired from different viewpoints.

Test	Ranking					
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
Lift.	T.Rnd	T.LR	**Li**	Th	Po	Wa
Point.	Th	Li	T.LR	Wa	T.Rnd	**Po**
Throw.	Po	**Th**	T.Rnd	Wa	T.LR	Li
T. LR	T.Rnd	**T.LR**	Li	Po	Th	Wa
T. Rnd	T.LR	**T.Rnd**	Li	Th	Po	Wa
Wav.	**Wa**	Th	Po	T.LR	T.Rnd	Li

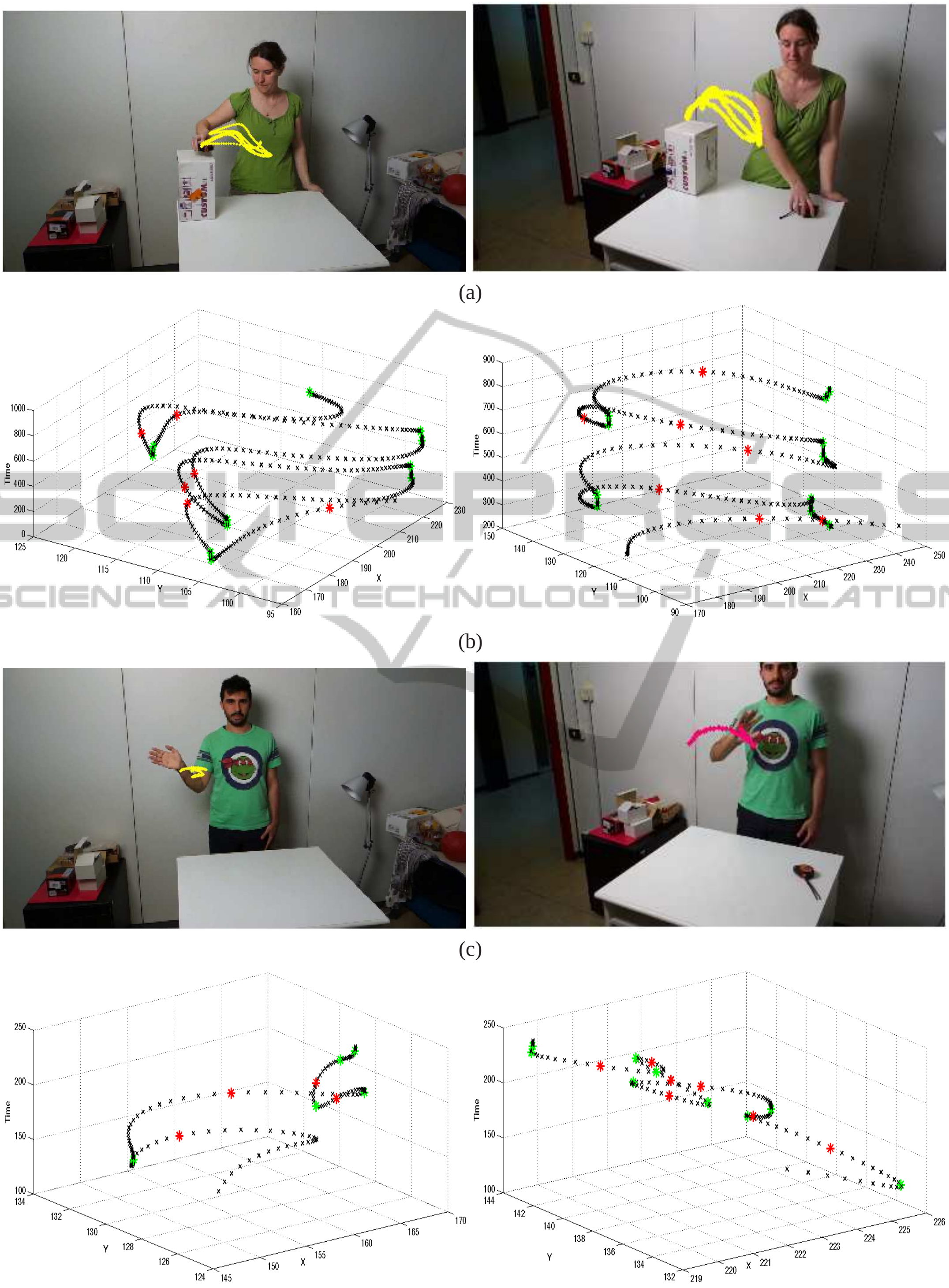
of trajectories related to different actions observed from the two viewpoints. On the same plot, we also report the local maxima of velocity magnitude and curvature. As a first thing, we may observe that the dynamic instants are tolerant to view-point changes. Furthermore, trajectories with diverse appearances in the image plane (e.g. with different lengths or spatial extents) present similar representations, showing that the dynamic instants are also tolerant to variations among the input data.

Action Recognition. We consider different experimental configurations of increasing complexity:

- Same subject same view
- Different subject same view
- Matching across different views.

For each configuration, we consider each time a video of test and match it with videos of all the available actions, then rank the similarities we obtained. Such rankings are reported in Tab. 1, 2 and 3, where actions are referred to as Li (Lifting), Po (Pointing), Th (Throwing), T.LR (Transporting left-right), T.Rnd (Transporting random), and Wa (Waving).
It is apparent how the increasing complexity of the configurations are reflected on the ranking results. If the matching performs accurately when considering videos of the same subject and from the same viewpoint, the results degrade already when the comparisons involve different subjects, even if observed from the same viewpoints. This is consequence of the movements subjectivity, that cause the presence of maybe subtle properties in the motion that fail to be captured from the basic representation we adopted.

Actions Type Affinity. What is interesting to be observed is that the comparison of the rankings of Tab. 2 and 3 suggests the presence of some equivalence classes between actions. To clarify this point, we perform an analysis of the overall similarities between actions and visualize the results in the similarity matrix of Fig. 6. Some remarks are in order. The first


(a)

(b)

(c)

(d)

Figure 5: Examples of dynamic instants extracted from two views and considering different actions: Lifting (Fig.5(a) and 5(b)) and Waving (Fig.5(c) and 5(d)). Local extrema of velocity magnitude (red) and curvature (green) are marked.

is that the computed similarities are high on average (all above 0.9) speaking in favor of the complexity of the problem. Second, some affinities between actions can be inferred. Waving appears as the most distinctive action among the considered set. The Transporting actions (both versions) are highly similar to each other. Moreover, they show affinities with Lifting. Pointing seems to share properties with Transport random (probably because of the variability in the movements direction), but also with Throwing. The latter, not influenced by the forces caused by the manipulation of objects, is thus more related to an intransitive action type. We further go on with the analysis by thresholding the similarity matrix with respect to the lowest value on the diagonal (i.e. the lowest similarity between two actions that *must* be similar). After that, we inferred the graph-like structure in Fig. 7, where the survived elements – that can be interpreted as affinities between the involved actions – are represented as arrows. It is easy to observe how the *transitive* actions (in green) formed a cluster, which is connected to an *intransitive* action (red) probably due to the affinities between the two dynamics. *Pointing* is also related to *Throwing*, that as above mentioned can be considered (in our instance) as an *intransitive* action.

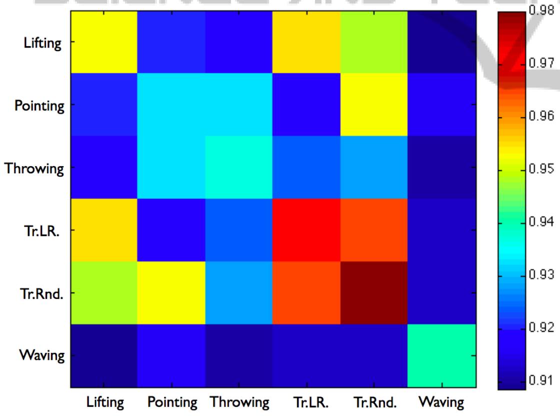


Figure 6: Average mutual similarity.

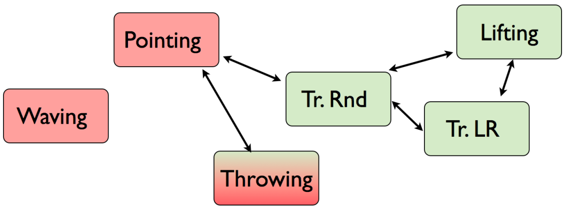


Figure 7: A visual sketch of the actions affinities inferred by mean of the analysis (in green *transitive* actions, in red *intransitive* actions). *Throwing* has both colors since in our instance we did not actually manipulate any object.

4 CONCLUSIONS

This work considered the problem of extracting knowledge about affinities between (classes of) actions across different views, with specific reference to the context of human-robot interaction. Starting from investigations about human motion perception and inspired from the seminal work (Rao et al., 2002), we considered a coarse motion description based on the use of histograms of curvature, which we used to match pairs of videos with a multi-level approach. We experimentally inferred a set of semantic connections that characterize actions groups across views.

Our observations, made with computational tools, confirm what observed directly in infants: they develop rather early the ability of grasping some aspects of the actions meaning, while it is likely that the capability of interpreting more specific actions properties is developed later. This sets the scene for the replicability on an artificial system – a robot in our case – of the early stages of the human developmental evolution, in which the interpretation of the observed motion refines more and more while strengthening the perceptual capabilities. In general an interactive robot needs to be able to autonomously understand where to focus its attention, for instance by perceiving the presence of motion, and biological motion in particular as an effect of a potential interacting agent. Starting from that, already by recognizing the class of actions of the partner (e.g. some kind of object manipulation) the robot could focus on the most relevant properties of the event (e.g. the manipulated object or the effects on the context). Finally, the understanding of the specific action and of its goal may guide the robot to the selection of an appropriate reaction (e.g. being prepared to receive an object from the user). Following this mainstream, our future investigations will be devoted to the development of a multi-level system for action recognition, in which the complexity of the computational model reflects the complexity of the task.

From the point of view of the vision task, future investigations will be also devoted to the design of models able to cope with different complexity of the scene (e.g. the presence of more than one moving agents).

REFERENCES

Aggarwal, J. and Ryoo, M. (2011). Human activity analysis: A review. *ACM Computing Surveys*.

Camaioni, L. (2004). The role of declarative pointing in developing a theory of mind. *Infancy*, 5:291–308.

Fanello, S. R., Gori, I., Metta, G., and Odone, F. (2013). Keep it simple and sparse: Real-time action recognition. *J. Mach. Learn. Res.*, 14(1):2617–2640.

Farah, M. J., Wilson, K. D., Drain, H. M., and Tanaka, J. R. (1995). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research*, 35(14):2089 – 2093.

Gong, D. and Medioni, G. (2011). Dynamic manifold warping for view invariant action recognition. *ICCV*, pages 571–578.

Goren, C. C., Sarty, M., and Wu, P. Y. K. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56(4):544–549.

Huang, C.-H., Yeh, Y.-R., and Wang, Y.-C. (2012a). Recognizing actions across cameras by exploring the correlated subspace. In *ECCV*, volume 7583 of *Lecture Notes in Computer Science*, pages 342–351.

Huang, K., Zhang, Y., and Tan, T. (2012b). A discriminative model of motion and cross ratio for view-invariant action recognition. *IEEE Transactions on Image Processing*, 21(4):2187–2197.

Junejo, I. N., Dexter, E., Laptev, I., and Prez, P. (2011). View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):172–185.

Kanakogi, Y. and Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nat.Commun.*, 2:341–.

Kuhlmeier, V. A., Troje, N. F., and Lee, V. (2010). Young infants detect the direction of biological motion in point-light displays. *Infancy*, 15(1):83–93.

Lewandowski, M., Makris, D., and Nebel, J.-C. (2010). View and style-independent action manifolds for human activity recognition. In *ECCV*, volume 6316 of *Lecture Notes in Computer Science*, pages 547–560.

Li, B., Camps, O. I., and Sznaier, M. (2012). Cross-view activity recognition using hankelets. In *CVPR*, pages 1362–1369.

Li, R. and Zickler, T. (2012). Discriminative virtual views for cross-view action recognition. In *CVPR*, pages 2855–2862.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110.

Mahbub, U., Imtiaz, H., Roy, T., Rahman, S., and Ahad, A. (2011). Action recognition from one example. *Pattern Recognition Letters*.

Malgireddy, M. R., Inwogu, I., and Govindaraju, V. (2012). A temporal bayesian model for classifying, detecting and localizing activities in video sequences. In *CVPRW*.

Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The icub humanoid robot: An open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, PerMIS '08, pages 50–56.

Noceti, N. and Odone, F. (2012). Learning common behaviors from large sets of unlabeled temporal series. *Image and Vision Computing*, 30(11):875 – 895.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *PAMI*, 12(7):629–639.

Rao, C., Yilmaz, A., and Shah, M. (2002). View-invariant representation and recognition of actions. *IJCV*, 50(2):203–226.

Sciutti, A., Noceti, N., Rea, F., Odone, F., Verri, A., and Sandini, G. (2014). The informative content of optical flow features of biological motion. In *ECVP*.

Shi, J. and Tomasi, C. (1994). Good features to track. In *CVPR*, pages 593 – 600.

Simion, F., Regolin, L., and Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, 105(2):809–813.

Troje, N. F. and Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a life detector? *Current Biology*, 16(8):821 – 824.

Wang, X., Ma, X., and Grimson, W. (2009). Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE transactions on pattern analysis and machine intelligence*, 31(3):539–555.

Welch, G. and Bishop, G. (1995). An introduction to the kalman filter. Technical report.

Wu, X. and Jia, Y. (2012). View-invariant action recognition using latent kernelized structural svm. In *ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 411–424.

Zheng, J. and Jiang, Z. (2013). Learning view-invariant sparse representations for cross-view action recognition. In *ICCV*, pages 3176–3183.

Zheng, J., Jiang, Z., Phillips, P. J., and Chellappa, R. (2012). Cross-view action recognition via a transferable dictionary pair. In *British Machine Vision Conference*, pages 1–11.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *ICPR*, volume 2, pages 28–31.