Investigating the Use of Space-Time Primitives to Understand Human Movements

Damiano Malafronte[1], Gaurvi Goyal[1], Alessia Vignolo[1,2], Francesca Odone[1], and Nicoletta Noceti[1](✉)

[1] Università degli Studi di Genova, Genova, Italy
{damiano.malafronte,gaurvi.goyal}@dibris.unige.it,
alessia.vignolo@iit.it, {francesca.odone,nicoletta.noceti}@unige.it
[2] Istituto Italiano di Tecnologia, Genova, Italy

Abstract. In this work we start investigating the use of appropriately learnt space-time primitives for modeling upper body human actions. As a study case we consider cooking activities which may undergo large intra class variations and are characterized by subtle details, observed by different view points. With a BoK procedure we quantize each video frame with respect to a dictionary of meaningful space-time primitives, then we derive time series that measure how the presence of different primitives evolves over time. The preliminary experiments we report are very encouraging on the discriminative power of the representation, also speaking in favor of the tolerance to view point changes.

Keywords: Spatio-temporal interest points · Motion primitives · Multi-view motion analysis · Multi-view action analysis · Shearlet transform

1 Introduction

Understanding human motion and its regularities is a key research goal of Human-Machine Interaction, with a potential to unlock more refined abilities – such as the anticipation of action goals – and thus the design of intelligent machines able to proficiently and effectively collaborate with humans [1,2].

In this ongoing work we are interested in investigating HMI functionalities, where a machine (e.g. a robot) observes a human performing tasks and learns how to discriminate among the ones characterized by different dynamic properties [3]. We consider upper body human action primitives taking place in a specific setting, cooking in our case. For the time being, we restrict our attention to the actor, and do not exploit any contextual information which could be derived, for instance, by the presence of a tool or an object.

Since some time we have assisted to a growing interest towards the so-called space-time key-points. From the pioneering work of Laptev [4], who proposed an extension to the space-time of corner points, soon followed by alternative and possibly richer approaches [5,6], we have appreciated the power of these key-points as low level building blocks for motion analysis and action recognition.

S. Battiato et al. (Eds.): ICIAP 2017, Part I, LNCS 10484, pp. 40–50, 2017.
https://doi.org/10.1007/978-3-319-68560-1_4

Space-time key points mark special points where the signal undergoes a significant variation both in space and time, and for this reason they are quite rare. They carry meaningful information in particular when we analyze distinctive dynamic events, but they may be not as effective with more subtle actions or gestures.

In this work, instead of retaining the sole information provided by these hand-crafted space-time key-points, we learn *ad hoc* space-time local primitives for a given (class of) action(s). Given a dynamic event, different meaningful local primitives can be observed and associated with an appropriate meaning in space and time [7]. To achieve this goal we follow and unsupervised approach and consider a signal representation based on Shearlets [8,9]. Shearlets emerge among multi-resolution models by their ability to efficiently capture anisotropic features, to detect singularities [10,11] and to be stable against noise and blurring [12–14]. The effectiveness of Shearlets is supported by a well-established mathematical theory and confirmed by a variety of applications to image processing [9,14,15].

We propose a pipeline to represent the space-time information embedded in an image sequence. First, from the $2\mathrm{D}+\mathrm{T}$ shearlet coefficients we represent a space-time neighborhood by appropriately encoding the signal behavior in space and time. Then, we learn a dictionary of space-time local primitives or atoms meaningful for a specific action set. To do so, we follow a BoK approach [16], applying a clustering procedure to all the space-time points of a training set of image frames. The whole procedure is carried out in an unsupervised way, in the sense we do not use labels describing specific image features. Finally, we represent a video sequence as a set of time series depicting the evolution of the primitives frequency over time.

In the preliminary results we report, we analyze this information and evaluate whether it is meaningful and stable to multiple repetitions of the same action and discriminative among different but similar actions. We also evaluate its robustness to view point variations and investigate the descriptive power of dictionaries learnt by different datasets. Instead of addressing view-invariance as a general property we focus on a set of different view points that describe typical observation points in human-human interaction (ego-view, frontal view, lateral view) as they are meaningful to a natural HMI.

2 Shearlet Theory: An Overview

Here we briefly review the construction of the discrete shearlet transform of a $2D+T$ signal f by adapting the approach given in [17] for 3D signals.

Denoted by L^2 the Hilbert space of square-integrable functions $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{C}$ with the usual scalar product $\langle f, f' \rangle$, the discrete shearlet transform $SH[f]$ of a signal $f \in L^2$ is the sequence of coefficients

$$SH[f](\ell, j, k, m) = \langle f, \Psi_{\ell,j,k,m} \rangle$$

where $\{\Psi_{\ell,j,k,m}\}$ is a family of filters parametrized by

1. A label $\ell = 0, \dots, 3$ of 4 regions or pyramids $\mathcal{P}_\ell$ in the frequency domain;
2. The scale parameter $j \in \mathbb{N}$;
3. The shearing vector $k = (k_1, k_2)$ where $k_1, k_2 = -\lceil 2^{j/2} \rceil, \dots, \lceil 2^{j/2} \rceil$;
4. The translation vector $m = (m_1, m_2, m_3) \in \mathbb{Z}^3$.

For $\ell = 0$ the filters, which do not depend on j and k, are

$$\Psi_{0,m}(x, y, t) = \varphi(x - cm_1)\varphi(y - cm_2)\varphi(t - cm_3), \tag{1}$$

where $c > 0$ is a step size and φ is a $1D$-scaling function. The system $\{\Psi_{0,m}\}_m$ takes care of the low frequency cube $\mathcal{P}_0 = \{(\xi_1, \xi_2, \xi_3) \in \widehat{\mathbb{R}}^3 \mid |\xi_1| \leq 1, |\xi_2| \leq 1, |\xi_3| \leq 1\}$.

For $\ell = 1$ the filters are defined in terms of translations and two linear transformations (parabolic dilations and shearings)

$$A_{1,j} = \begin{pmatrix} 2^j & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix} \qquad S_{1,k} = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{so that}$$

$$\Psi_{1,j,k,m}(x, y, t) = 2^j \psi_1 \left(S_{1,k} A_{1,j} \left(\begin{smallmatrix} x \\ y \\ t \end{smallmatrix}\right) - \left(\begin{smallmatrix} cm_1 \\ \hat{c}m_2 \\ \hat{c}m_3 \end{smallmatrix}\right) \right), \tag{2}$$

where c is as in (1) and $\widehat{c} > 0$ is another step size (in the rest of the paper we assume that $c = \hat{c} = 1$ for sake of simplicity). The system $\{\Psi_{1,j,k,m}\}$ takes care of the high frequencies in the pyramid along the x-axis: $\mathcal{P}_1 = \{(\xi_1, \xi_2, \xi_3) \in \widehat{\mathbb{R}}^3 \mid |\xi_1| \geq 1, |\frac{\xi_2}{\xi_1}| \leq 1, |\frac{\xi_3}{\xi_1}| \leq 1\}$. For $\ell = 2, 3$ we have a similar definition by interchanging the role of x and y (for $\ell = 2$) and of x and t (for $\ell = 3$).

Our algorithm is based on a nice property that allows us to associate with any shearing vector $k = (k_1, k_2)$ a direction (without orientation) parametrized by two angles, namely *latitude* and *longitude*, given by

$$(\cos\alpha \cos\beta, \cos\alpha \sin\beta, \sin\alpha) \qquad \alpha, \beta \in [-\frac{\pi}{2}, \frac{\pi}{2}]. \tag{3}$$

The correspondence depends on ℓ and, for the first pyramid, it is given by

$$\tan\alpha = \frac{2^{-j/2} k_2}{\sqrt{1 + 2^{-j} k_1^2}} \qquad \tan\beta = 2^{-j/2} k_1 \quad \alpha, \beta \in [-\frac{\pi}{4}, \frac{\pi}{4}].$$

The fact that Shearlets are sensitive to orientations allows us to discriminate among spatial-temporal features of different kinds [7,18].

3 Building Dictionaries of Space-Time Primitives

1 - Space-Time Point Representation (Fig. 1). We start by considering a point $\hat{m}$ for the fixed scale $\hat{j}$ and the subset of shearings encoding different directions: $\mathbf{K} = \left\{ k = (k_1, k_2) \mid k_1, k_2 = -\lceil 2^{\hat{j}/2} \rceil, \dots, \lceil 2^{\hat{j}/2} \rceil \right\}$. We perform the following steps:

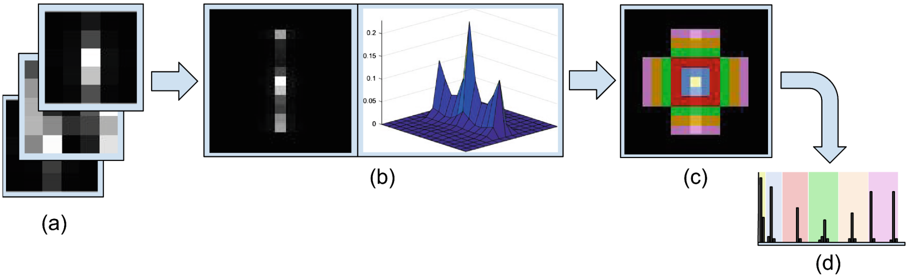


Fig. 1. $2D+T$ point representation: (a) Matrices $C_1(r,c)$, $C_2(r,c)$ and $C_3(r,c)$; (b) Object **C** both in gray-levels and 3D visualization; (c) Coefficients grouping; (d) The obtained representation **D**.

Figure 1a. We reorganize the information provided by $SH[f](\ell,\hat{\jmath},k,\hat{m})$ in three $M \times M$ matrices, each one associated with a pyramid ℓ, where each entry is related to a specific shearing: $C_\ell(r,c) = SH[f](\ell,\hat{\jmath},k_{rc},\hat{m})$ with $\ell = 1,2,3$, where r and c, are discrete versions of k_1 and k_2.

Figure 1b. We merge the three matrices in a single one. The obtained overall representation **C** is centered on k_{max}, the shearing corresponding to the coefficient with the maximum value in the set $SH[f](\ell,\hat{\jmath},k,\hat{m})$, with $\ell \in \{1,2,3\}$ and $k \in \mathbf{K}$. The matrix **C** models how the shearlet coefficients vary in a neighborhood of the direction where there is the maximum variation, and it is built in a way so that the distance of every entry of **C** with respect to the center is proportional to the distance of the corresponding angles (as defined in (3)) from the angles associated with k_{max}. Different kinds of spatio-temporal elements can be associated with different kinds of local variations in **C** (see for instance Fig. 6).

Figure 1c. We now compute a compact rotation-invariant representation for point $\hat{m}$. We group the available shearings in subsets $\bar{s}_i$, according to the following rule: $\bar{s}_0 = \{k_{max}\}$ and $\bar{s}_i$ will contain the shearings in the *i-th* ring of values from k_{max} in **C**. We extract the values corresponding to the coefficients for $\bar{s}_1$ (by looking at the 8-neighborhood of k_{max}), then we consider the adjacent outer ring (that is, the 24-neighborhood without its 8-neighborhood) to have the coefficients corresponding to $\bar{s}_2$, and so on.

Figure 1d. We build a vector containing the values of the coefficients corresponding to each set: $\mathbf{D}(\hat{m}) = \mathit{coeff}_{\bar{s}_0} {}^\frown \mathit{coeff}_{\bar{s}_1} {}^\frown \mathit{coeff}_{\bar{s}_2} {}^\frown \ldots$; $\mathit{coeff}_{\bar{s}_i}$ is the set of coefficients associated with each shearings subset $\bar{s}_i$:

$$\begin{aligned} \mathit{coeff}_{\bar{s}_0} &= SH[f](\ell_{k_{max}},\hat{\jmath},k_{max},\hat{m}) \\ \mathit{coeff}_{\bar{s}_i} &= \left\{ SH[f](\ell_{\bar{s}_i},\hat{\jmath},k_{\bar{s}_i},\hat{m}), k_{\bar{s}_i} \in \bar{s}_i \right\}, \end{aligned}$$

where $\ell_{k_{max}}$ is the pyramid associated with the shearing k_{max} and where $\ell_{\bar{s}_i}$ represents the pyramid associated to each shearing $k_{\bar{s}_i}$.

2 - Learning a Dictionary of Space-Time Primitives (Figure 2).

Figure 2a. This phase considers a set of meaningful frames in a (set of) sequence(s). The frames are chosen automatically through a key-point detection process [18]. We select the N_f frames with the highest number of interest points and we assume that these are the most representative of an action event.

Figure 2b. We represent each point $\hat{m}$ of every selected frame by means of $\mathbf{D}(\hat{m})$, for a fixed scale $\hat{j}$. On each frame, we apply K-means and obtain a set of K cluster centroids, which we use as space-time primitives or atoms.

Figure 2c. We re-apply K-means on all the previously obtained atoms [7]. We end up with a dictionary $\mathcal{D}$ of N_a space-time primitives.

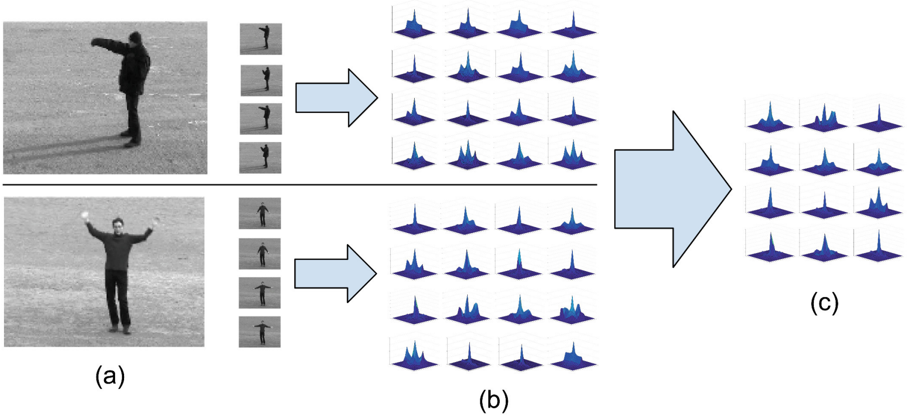


Fig. 2. Learning the dictionary. (a) Automatic selection of meaningful frames from the training set; (b) Atoms learnt by each sequence; (c) Dictionary summarization on the whole training set.

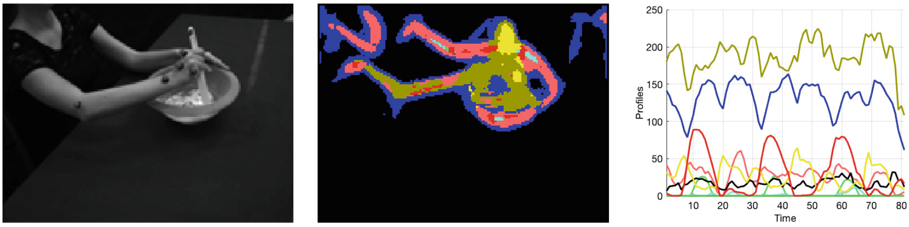
3 - Encoding a Video Sequence with Respect to a Dictionary (Figure 3). We now consider a sequence V of a given action.

Figure 3b. For each image frame $I_t \in V$ we follow a BoK approach and quantize points of I_t w.r.t the dictionary atoms, obtaining F_i^t frequency values (how many points in frame I_t can be associated with the $i-th$ atom).

Figure 3c. We filter out still primitives that are not useful to our purpose. To do this, we consider a point-wise index which we call *dynamism measure* (DM):

$$\mathrm{DM}[\hat{m}] = SH[f](\ell_{k_{max}}, j, k_{max}, \hat{m}) \cdot cos(\varTheta_{k_{max}}, \boldsymbol{n}) \tag{4}$$

where for a given point $\hat{m}$ we consider the value corresponding to its maximum shearlet coefficient and its associated shearing parameter k_{max}; $\varTheta_{k_{max}}$ is the associated direction obtained using (3) and $\boldsymbol{n}$ is the normal vector to the xy plane in our signal (i.e. aligned with the temporal axis). To discard still patterns we consider only the values of $\mathrm{DM}[\hat{m}]$ which are above a given threshold τ. The angle $\Theta_{k_{max}}$ tells us whether a point belongs to a spatio-temporal structure which is moving or not[1], while the $SH[f](\ell_{k_{max}}, j, k_{max}, \hat{m})$ factor helps us to consider only points representing a *strong* spatio-temporal change. Finally, we compute temporal sequences of frequency values across time, obtaining N_a time series or profiles $\{P_j\}_{j=1}^{N_a}$, which summarize the content of the video sequence.



(a) (b) (c)

Fig. 3. Action encoding: (a) A sample frame; (b) The quantization w.r.t. the dictionary atoms; (c) Examples of temporal profiles (see text for details).

4 Experimental Analysis

4.1 Dataset and Experimental Protocol

The data we consider are drawn from a larger dataset of cooking actions that we will soon release to the research community. We have used three identical high resolution IP cameras, mounted on three tripods so that in all acquisitions we have a still uniform background and moving foreground objects. Figure 4 shows the setup and example video frames. The dataset includes repetitions of the same action observed from three different viewpoints: a frontal view (A), a lateral view (B), and an egocentric view, obtained by a camera mounted slightly above the subject's head (C). No specific constraints have been imposed to the volunteer.

For this preliminary analysis we are considering a subset of 3 actions. For each action and each view we consider 3 action instances. In the following experiments we consider dictionaries learnt from *Eating* actions only. For the detection phase (see [18]), we fix the number of selected frames N_f to 4 and consider only shearlet coefficients at scale 2. For the dictionary learning phase, the number of centroids per frame is $K = 8$, and the final dictionary size is $N_a = 12$.

[1] Points belonging to still spatio-temporal structure spawn surfaces over time, and the normal vector $\Theta_{k_{max}}$ for those points will belong to the xy plane, bringing the value for $cos(\Theta_{k_{max}}, \boldsymbol{n})$ to be 0.

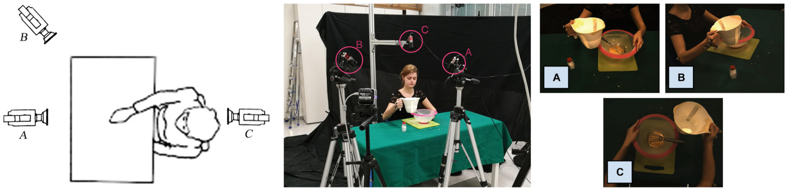


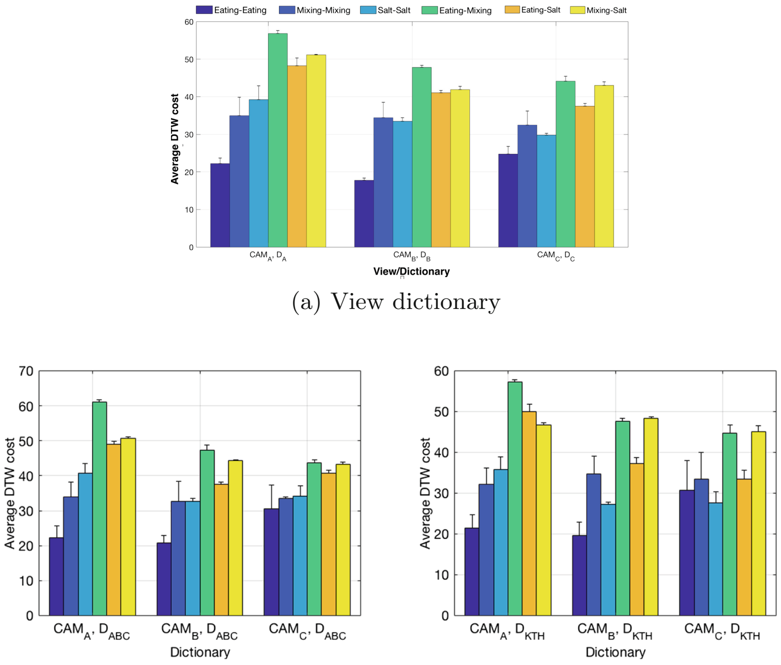
Fig. 4. Acquisition setup

We evaluate the dissimilarity between action pairs by means of Dynamic Time Warping (DTW). Given two videos V^1 and V^2 depicting a certain action instance and described by two sets of temporal profiles $P^1 = \{P_i^1\}_{i=1}^{N_a}$ and $P^2 = \{P_i^2\}_{i=1}^{N_a}$ then $Dis(V^1, V^2) = \text{avg}_{i=1}^{N_a} DTW(P_i^1, P_i^2)$. Z-normalization is applied to the temporal profiles before computing the dissimilarity.

4.2 Preliminary Investigation

1. How informative are the learnt space-time dictionaries to discriminate among different actions of the same kind? In this experiment we consider comparisons between actions observed from a given viewpoint, described according to a dictionary obtained from the same view: we refer to such dictionaries as D_A, D_B, and D_C. In Fig. 5a we show the average DTW cost in aligning the instances of the action classes. We observe that on average the comparisons of actions from the same class have a lower cost. Among the 3, CAM_C appears to be the most challenging viewpoint. We may notice that *Eating* action is the best performing, as dictionaries are built on eating examples. At the same time we observe a good generalization to other actions.

2. What is the relationship between different dictionaries learnt from different viewpoint data? Is there any benefit in learning dictionaries from different views? To answer this question, we compare dictionaries specific to different views, and observe they encode similar spatio-temporal primitives. We build a dissimilarity matrix collecting the Euclidean distances between atoms of the two dictionaries. The atoms are then matched using the Hungarian algorithm, and their contributions are sorted in the dissimilarity matrix accordingly. As a consequence, on the main diagonal we may find agglomerations of atoms belonging to different dictionaries but encoding the same kind of spatio-temporal information. Figure 6 shows an example where dictionaries referring to CAM_A and CAM_B are considered, and where we highlighted groups of atoms carrying similar information. At the top of the diagonal a group of 3 atoms (Fig. 6a) describe moving edge-like structures, which correspond to surface in the space-time domain. Similarly, the primitives in Fig. 6b and c represent corner-like structures with a different amount of dynamic variations in the direction around the principal one.



(a) View dictionary

(b) Combined dictionary

(c) KTH dictionary

Fig. 5. Average DTW cost obtained when comparing actions of the same view using different dictionaries.

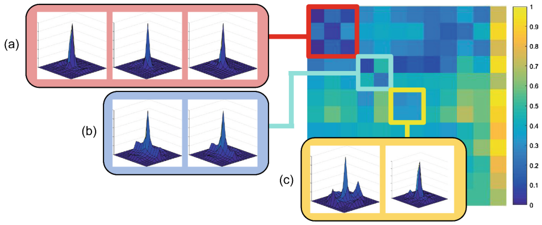
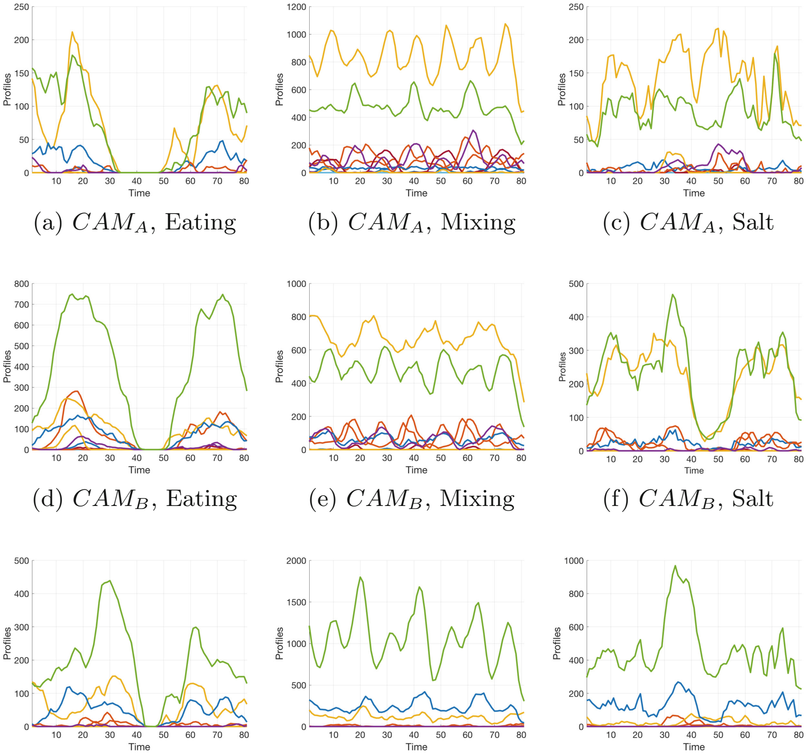


Fig. 6. An example of dissimilarity matrix between atoms of two different dictionaries (from CAM_A and CAM_B), with a selection of prototypes encoding different dynamic properties of the signal.

As we observe a large overlap between different dictionaries, we also consider the benefits of learning a joint dictionary from the 3 views, as this choice would simplify inter-view comparisons. Figure 5b shows how stable the performance is when adopting D_{ABC} for all the data.

3. To what extent the space-time representation is view-invariant? Figure 7 provides a first qualitative answer to the question. The plots represent the average profiles of all actions instances. *Eating* is characterized by the high-



(a) CAM_A, Eating (b) CAM_A, Mixing (c) CAM_A, Salt

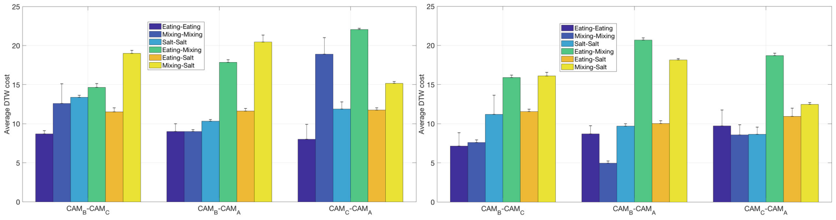
(d) CAM_B, Eating (e) CAM_B, Mixing (f) CAM_B, Salt

(g) CAM_C, Eating (h) CAM_C, Mixing (i) CAM_C, Salt

Fig. 7. Average temporal profiles of different action instances. Each row corresponds to a view (CAM_A, CAM_B, CAM_C), while each column refers to an action (Eating, Mixing, Salt). The dictionary D_{ABC} is employed.

est stability across views, while *Mixing* presents some differences in CAM_C with respect to the other two views. This may be explained with the fact the action is performed following a quasi-planar shape on the table, favouring a clear and regular apparent motion from the top view. *Salt* is a less constrained action characterized by a higher degree of instability over time and across views. Figure 8a reports the average DTW costs obtained from pairs of views. On the left (D_{ABC}) we confirm *Eating* is stable across views, while a higher intra-class variability is associated with *Mixing*. We also notice a similarity between *Eating* and *Salt*. A visual inspection of the corresponding profiles in Fig. 7 confirms the presence of common temporal patterns.

We observe that the different temporal profiles are characterized by an uneven amount of stability. This suggests that a selection of the profiles to be used in the comparison may be of benefit. This aspect is currently under investigation, as a proof of concept, in Fig. 8b we consider only one profile, the green one in Fig. 7. An improvement on the results may be appreciated.



(a) (b)

Fig. 8. Comparison between descriptions from different views.

4. Is it really useful to learn an ad hoc dictionary for a given set of data? As a final investigation, we reason on the necessity of using data of the considered scenario. To this purpose we consider an unrelated benchmark (KTH [19]) showing full body actions. Figure 5c shows the results obtained in this case. We notice a small degradation, but the overall performance is still acceptable. This speaks in favor of the potential of our space-time primitives to transfer knowledge between different settings.

5 Discussion

We presented an ongoing work on representing actions through space-time primitives learnt from data. The preliminary results on a small subset of data include useful insights on how to proceed: the representation is rich and incorporates not only space-time corners but also other local structures with a significant dynamic information; the learnt atoms are quite stable across views, with a strong discriminative power. The action representation is again quite stable across views, even if some actions seem to be intrinsically view-variant, and some views are more meaningful than others. Representations obtained from front and lateral views are very closely related, as expected.

Two main aspect are currently under investigation: *(i)* Capturing the temporal cross-correlation between different primitives, especially across views and *(ii)* Devising an action recognition module based on the proposed representation.

References

1. Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: the origins of cultural cognition. Behav. Brain Sci. **28**(05), 675–691 (2005)
2. Elsner, C., Falck-Ytter, T., Gredebäck, G.: Humans anticipate the goal of other people's point-light actions. Front. Psychol. **3** (2012)
3. Vignolo, A., Rea, F., Noceti, N., Sciutti, A., Odone, F., Sandini, G.: Biological movement detector enhances the attentive skills of humanoid robot iCub. In: IEEE-RAS (Humanoids), pp. 338–344 (2016)

4. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2), 107–123 (2005)
5. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
6. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88688-4_48
7. Malafronte, D., Odone, F., De Vito, E.: Local spatio-temporal representation using the 3D shearlet transform. In: SAMPTA (2017)
8. Labate, D., Lim, W.-Q., Kutyniok, G., Weiss, G.: Sparse multidimensional representation using shearlets. In: Optics & Photonics (2005)
9. Kutyniok, G., Labate, D.: Shearlets. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York (2012). doi:10.1007/978-0-8176-8316-0
10. Guo, K., Labate, D., Lim, W.-Q.: Edge analysis and identification using the continuous shearlet transform. Appl. Comput. Harmon. Anal. **27**(1), 24–46 (2009)
11. Kutyniok, G., Petersen, P.: Classification of edges using compactly supported shearlets. Appl. Comput. Harmon. Anal. **42**(2), 245–293 (2015)
12. Easley, G.R., Labate, D., Colonna, F.: Shearlet-based total variation diffusion for denoising. TIP **18**(2), 260–268 (2009)
13. Chen, Z., Hao, X., Sun, Z.: Image denoising in shearlet domain by adaptive thresholding. J. Inf. Comput. Sci. **10**(12), 3741–3749 (2013)
14. Duval-Poo, M.A., Noceti, N., Odone, F., De Vito, E.: Scale invariant and noise robust interest points with shearlets. IEEE Trans. Image Process. **26**(6), 2853–2867 (2017)
15. Duval-Poo, M.A., Odone, F., De Vito, E.: Edges and corners with shearlets. IEEE Trans. Image Process. **24**(11), 3768–3780 (2015)
16. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV-W, vol. 1, no. 1–22 (2004)
17. Kutyniok, G., Lim, W., Reisenhofer, R.: Shearlab 3D: faithful digital shearlet transforms based on compactly supported shearlets. ACM Trans. Math. Softw. **42**, 5:1–5:42 (2016)
18. Malafronte, D., Odone, F., De Vito, E.: Detecting spatio-temporally interest points using the shearlet transform. In: Alexandre, L.A., Salvador Sánchez, J., Rodrigues, J.M.F. (eds.) IbPRIA 2017. LNCS, vol. 10255, pp. 501–510. Springer, Cham (2017). doi:10.1007/978-3-319-58838-4_55
19. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR, vol. 3 (2004)